eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Multilingual Sentiment Analysis in Social Media

Iñaki San Vicente Roncal

PhD Thesis

informatika fakultatea

facultad de informática

Donostia, January 2019

Lengoaia eta Sistema Informatikoak Saila

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Informatika Fakultatea

Multilingual Sentiment Analysis in Social Media

Thesis written by Iñaki San Vicente Roncal under German Rigau and Rodrigo Agerri's guidance,
presented to obtain the title of Doctor in Computer Science in the University of the Basque Country

Donostia, January 2019

Acknowledgements

First of all, I would like to thank my colleague at Elhuyar Xabier Saralegi, who has been my supervisor and partner for the longest part of my research career including most of the experiments carried out in the Sentiment Analysis field. Without his participation and guidance this work would not have been possible.

I would also like to thank my advisors at University, Dr. Rodrigo Agerri and Dr. German Rigau, who have greatly helped me and guided me through the process of writing this manuscript.

I am grateful to all my colleagues in Elhuyar. My group has provided me with all the resources that I have needed for my work and they have never hesitated to give me a helping hand. I hope to continue this collaboration in many future projects.

I would like to show my gratitude to the IXA group. Their interest in my research was really encouraging, and my one year stay at the group was a great impulse towards this work.

A mention goes to the projects and institutions that have partially supported this research over the years: Industry Department of the Basque Government under grants IE12-333 (*Ber2tek*), SA-2012/00180 (*BOM2*), KK-2015/00098 (*Elkarola*) and KK-2017/00043 (*BerbaOla*); Spanish Government MICINN projects *Skater* (Grant No. TIN2012-38584-C06-01), *Tacardi* (Grant No. TIN2012-38523-C02-01) and *Tuner* (Grant No. (TIN2015-65308-C5-1-R); Behagunea project (San Sebastian European Capital of Culture 2016); and *OpeNER* European FP7 project (Grant No. 296451).

Last but not least, this work would have not been possible without the support of my family, especially Garazi and Oinatz, who had and still have to bear with my long working hours so many times. Thank you for just everything.

Contents

List of Figures

List of Tables

INTRODUCTION

CHAPTER 1

Introduction

This thesis is the result of a research journey started in 2011 and which still continues nowadays. Given that no work had been done at the time in Sentiment Analysis (SA) for the Basque language, the initial goal was to provide the necessary resources to perform SA for this language. The lack of even the most basic resources such as polarity lexicons or sentiment annotated corpora, made us look for approaches which were both adequate for less resourced languages and also affordable in terms of creation costs. That was also one of the main reasons to turn our attention to social media, as a source of user generated opinions that could be harvested in a cheap manner. This in turn brought its own challenges, such as language identification and microtext normalization.

This dissertation is the account of this journey, from the first experiments for the construction of subjective vocabulary lexicons to a full multilingual polarity classification system, including language normalization and both supervised and unsupervised classifiers. The thesis is organized as a collection of papers published in the aforementioned research framework.

This introduction is organized as follows. Next section further explains the motives behind this thesis. Section 1.2 describes our research framework and presents some concepts to introduce the reader in the notion of Sentiment Analysis and the particularities of applying it to a social media environment. After that, section 1.3 describes the main goals of our research. Finally, we describe the organization of the rest of the document in section 1.5.

1.1 Motivation

Why develop Sentiment Analysis for Basque in the first place? Research efforts in this field have exponentially increased in the last years, due to its applicability in areas such as Technological Surveillance/Competitive Intelligence, marketing or reputation management. The Internet has become a very rich source of user-generated information. Consumers' opinions are now public and accessible to everyone in the Web. Organizations are increasingly turning their eyes to this source in order to obtain global feedback on their activities and products. Examples of that are stock market prediction (Bollen et al., 2010; Oliveira et al., 2017), polling estimation (O'Connor et al., 2010; Ceron et al., 2015) or crisis events analysis (Pope and Griffith, 2016; Shaikh et al., 2017; Öztürk and Ayvaz, 2018). In the global world such information is multilingual, and so it is paramount to be able to harvest and process data in several languages.

But, is there enough user generated content to extract opinions from Basque texts? Is it worth the effort? If we look at opinion related websites there is almost no activity in Basque. Major specialized websites like TripAdvisor, Amazon, etc. do not have any content in Basque. There are very few specialized review sites, e.g., Armiarma (literature) or zinea.eus (movies). Even the best known Basque digital news media (Berria.eus, Sustatu.eus, Zuzeu.eus) do not have very active comment sections.

However, we find a very different scenario when we turn our attention to social media. "Euskararen adierazle sistema" (EAS), measures the digital health of the Basque language, among others. The most recent data (2016) reports that 25.7% of the population in the Basque Country (including Navarre and French Basque provinces) has activity in Basque in social media[1]. This number rises to 33.6% if we restrict our analysis to from a 16-50 year range which accounts for 70 to 80% of the users in Twitter. The last EAS study reports that up to a 15% of the content produced in Twitter by Basque Country inhabitants is done in Basque (december 2017). Umap[2], a website dedicated to track the Basque activity on Twitter, reports 2.5-2.8 million tweets per year, written in Basque[3][4].

[1] `http://www.euskadi.eus/web01-apeusadi/eu/eusadierazle/graficosV1.apl?~idioma=e&indicador=83`

[2] `www.umap.eus`

[3] `https://umap.eus/artxiboa`

[4] `https://www.codesyntax.com/eu/bloga/twitterreko%2Deuskarazko-jarduna%2D2016ko%2Dlaburpen%2Dtxostena%2Dumap`

The monitoring processes carried out with Talaia (San Vicente et al., 2019), the social media monitor developed as a result of this thesis, also prove that the number of opinions retrieved require an automatized analysis. For example, the monitoring of the Basque election campaign in September 2016[5] analysed an average amount of 2,500 mentions per day, 1,000 of them in Basque.

Would it be enough for a real world case scenario to be able to extract opinions exclusively in Basque? The aforementioned figures show the socio-linguistic reality of the Basque language, as it is for many less resourced languages. Basque coexists with both Spanish and French (also with English to a lesser extent), and most of its speakers are bilingual. Considering these issues, monitoring opinions with respect to a topic only in Basque would disregard a great number of opinions leading to an important coverage loss. Hence, the need to work on multilingual systems for analysis of opinions.

In summary, there is indeed a need for Basque Sentiment Analysis, particularly from social media sources, since they are by far the most active channel were users express their opinions in Basque. Furthermore, given that users talk about anything in social media, we can eventually crawl data for any domain from a single source, unlike specialized review sites. We have to bear in mind however that such source represents a particular genre of its own, and thus, that we will need to adapt our SA systems accordingly.

1.2 Research Framework

Sentiment Analysis is the sub-field of Natural Language Processing (NLP) that studies people's opinions, sentiments, and attitudes towards products, organizations, entities or topics. Although several complex emotion categorization models have been proposed in the literature (Russell, 1980; Ekman et al., 1987; Parrott, 2001; Plutchik, 2001; Cambria et al., 2010) most of the Sentiment Analysis community has assumed a simpler categorization consisting of two variables: subjectivity and polarity. A text is said to be **subjective** if it conveys an opinion, and objective otherwise. We understand **polarity classification** as the task of telling whether a piece of text (document, sentence, phrase or term) expresses a sentiment. This classification may be binary *[positive,negative]* or in a scale, e.g. *[positive—neutral—negative], [0..5].* Many researchers limit the polarity classification task to opinionated text, and treat objective statements as neutral. We should however have in mind that this is a

[5] `http://talaia.elhuyar.eus/demo_eae2016`

simplification, because objective facts (e.g. "My father died" - *negative*) can also bear sentiments, and thus be worth of our attention (Liu, 2012).

A second aspect would be to define what is the level of analysis we are interested in. The lowest level would be to spot sentiment bearing words (Stone et al., 1966) and expressions (Wilson et al., 2005). This is no trivial task since word polarity may be ambiguous or dependent on the context (see Example 1). We call sentiment or polarity lexicon a compiled list of those words and their polarities. Sentiment lexicons are basic resources for polarity classification.

Example 1
"Gure salmentek ***behera egin*** *dute"*[6] {*negative*} *vs. "Langabeziak* ***behera egin*** *du"*[7] {*positive*}

Earlier attempts to computationally assess sentiment in text were based on document classification (Pang et al., 2002; Turney, 2002; Wiebe et al., 2001; Hu and Liu, 2004a). Because many of them relied on the presence of polar words and/or co-occurrence statistics, a document guarantees to have a fair amount of clues. Also, some sources such as movie reviews provided ready-to-use document level annotations, making it possible to develop the first supervised systems (Pang et al., 2002).

Example 2
"Family hotel. Age is showing. ***Great*** *[staff]." A value hotel for sure with [rooms] that are average, however some* ***nice*** *touchs like the [coffee station] downstairs and the* ***free*** *[brownies] in the evening.* ***Great*** *[staff],* ***super friendly****. Special thanks to [Camilla] who was* ***very helpful and forgiving****, When we returned our damaged umbrella.* [8]

However, document level polarity classification is often not enough. Market analysis for example requires a more fine-grained analysis. Let's see the hotel review in Example 2. The overall sentiment may be average, but the hotel would profit for a more detailed information such as:

- Sentiment score towards staff is very positive.

[6]English translation: Our sales are going down.
[7]English translation: The unemployment rate is going down.
[8]Legend: Negative terms or phrases are underlined, positive ones are written in bold. Opinion targets are surrounded by square brackets.

- The building needs renovation.
- The worker called Camilla is very positively regarded.

Aspect Based Sentiment Analysis (ABSA) is the subfield that addresses this task (Nakov et al., 2013). Extracting this kind of information consists of several subtasks, such as detecting the holder (who says it), the expression (the words in text conveying the opinionated content) and the target of the opinion (what is the holder talking about). Targets talking about especific characteristics of the domain under analysis are called aspects. E.g., In example 2 opinion targets 'staff' or 'Camilla' would refer to the 'staff' aspect of a hotel. Such analysis may involve NLP tasks such as Named Entity Recognition and coreference resolution, apart from the polarity classification.

There is also the problem of the point of view. Let's look at the example 3. This is an objective sentence. However, if we were analysing the reputation of Osasuna football team, we should regard this sentence as positive. Similarly it would be negative referring to Valladolid football team.

Example 3
"Osasunak 4-2 irabazi zuen Valladoliden aurka".[9]

The work on this thesis is focused mostly on document and sentence level polarity classification because messages from social media are short documents and often with a single sentence. Nonetheless, ABSA will also be addressed to a certain extent.

As mentioned before, computational approaches to Sentiment Analysis use data from online review sites, such as Amazon, Yelp, TripAdvisor, etc. Such sources allow to build large datasets with document level annotations. They have however two main problems: i) They are domain specific, and few domains have this kind of specialized sites (hotels, restaurants, consumables, movies) ii) only global languages are used in this kind of sites. In consequence, it is very difficult to use such sources for less resourced languages. Even if we could gather a minimum size dataset out of those sources, it would make no sense to build a domain oriented system for a domain where little content is generated. Even so, some of the experiments carried out during this research do use this kind of sources, mainly for comparison with well established benchmarks. Thus, we focus on social media as an information source, and specifically on Twitter.

What are advantages and problems of this choice? The research community has worked extensively for some years now with Twitter data. On the one hand, Twitter

[9]English translation: Osasuna won 4-2 against Valladolid.

has an open policy with respect to sharing their data, including developer APIs, which makes it relatively easy to access the data. On the other, the large volume of messages and their real time nature, makes Twitter a very attractive information source for many tasks of predictive analysis, including SA.

Twitter however has also its particularities, which makes it more challenging to work with, especially from an NLP point of view. Twitter language is a genre of its own. First of all, the well-known 140 character limit[10] means that we will be dealing with short messages most of the time. Due to this limit, a new language has evolved in Twitter, including ungrammatical sentences, short messages, non-standard language, emojis, and other phenomena. Thus, traditional NLP tools need to be adapted to this new language (Wei et al., 2011), which is a challenge in itself. Example 4 presents a tweet that needs standardization followed by the closest correction proposal.

Example 4
"Loo Exoo Maazooo dee Menooss Puuff :(" →
"Lo hecho mazo de menos Puff :("[11]

In order to apply any NLP tool chain for analysis, a minimum requirement is to know the language of the message. Language identification is however still an open issue when analysing social media data.

Firstly, we find ourselves in a big-data environment, where the presence of less resourced languages is insignificant compared to others. At the begining of this thesis there was no support for Basque in social media. As of 2018, Twitter offers language identification for 60+ languages[12], including Basque. Facebook also reports being able to recognize 170+ languages, Basque among them[13]. Other social media are not as supportive [14]. Social media services usually identify major languages only. Other languages need to rely in their own language identification strategies. The problem accentuates for languages which are close to others, or have a social reality were they are mixed with a major language.

[10]The limit was expanded to 280 characters in November 2017, although Twitter reports that the average Tweet length still remains below 50 characters. In any case, the experiments presented in this dissertation were carried out over datasets built with the 140 limit in effect.
`https://www.theverge.com/2018/2/8/16990308/twitter-280-character-tweet-length`
`https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html`

[11]English translation: I miss him so much :(

[12]`https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-performance.html`

[13]`https://fasttext.cc/blog/2017/10/02/blog-post.html`

[14]`http://behatokia.eus/EU/albisteak/Interneten_ere_euskaraz_bizi_nahi_dugu`

Secondly, language identification in Twitter is affected by the aforementioned language particularities and also by the problem of mixed language content or code switching, as illustrated by Example 5. This thesis addresses those issues and explores the case of the languages coexisting in the Iberian peninsula.

Example 5
"Kaixo, acabo de hacer la azterketa de gizarte. Fatal atera zait! :(" [15] [16]

1.3 Main Goals

As mentioned at the beginning of this chapter, the ultimate goal was to develop a Sentiment Analysis system for Basque. However, because of the socio-linguistic reality of this language a tool providing only analysis for Basque would not suffice for a real world application. Thus, we set out to develop a multilingual system, which would ideally include Basque and the major languages coexisting with Basque: Spanish, French and English to a lesser extent.

At the moment of starting this thesis, machine learning methods had already become the more widely used approach over unsupervised systems that try to model complex linguistic phenomena by means of rules and polarity lexicons. That meant we needed sentiment annotated data to train our algorithms. Much work had been done for major languages, but less resourced languages such as Basque suffered (and they still do) from a lack of resources, both with respect to supervised (annotated training data) and unsupervised approaches (lexicons, rules).

This thesis addresses the problem of creating such resources, and proposes several cost and resource effective methods, with the final objective of constructing a multilingual Sentiment Analysis system that will be flexible enough to include non-major languages. Social media will be used as main data source, specifically Twitter, requiring some particular preprocessing. Thus, the following objectives were defined:

- **Analysing methods for creating Polarity lexicons**, comparing various approaches (Dictionary-based, corpus-based, manual methods), suitable for less resourced languages. As a result, we proposed a robust, multilingual approach which was deployed to create the first polarity lexicons for Basque.

[15] English translation: Hi, I just finished the exam of Social Studies class. I dit it awfully! :(

[16] Legend: underlined words are written in Basque, the rest in Spanish. At sentence level, the first one may be classified as Spanish, including Spanish grammar structure. The second one would be classified as Basque.

- **Analysis of social media**: As it was previously mentioned, tweets pose several challenges in order to understand and extract opinions from such messages. More specifically, two main issues needed to be addressed:
  - Language identification for less resourced languages.
  - Analysing methods for processing non standard language, emojis, etc. Much information which is relevant for SA (e.g. emojis often express the mood of the author) is lost if such elements are ignored. We will also address rule based systems and language model-based probabilistic approaches. Also the integration of such normalization into a SA system will be addressed.
- **Multilingual Sentiment Analysis system**: Research the state of the art in polarity classification, and to develop a supervised classifier that is tested against well known social media benchmarks. Creating training data for non major languages is also a primary goal putting special attention on how to crawl data from social media, general or domain specific.

1.4 Main contributions

As we have seen in the previous sections, analysing sentiment in a social media environment requires work on several NLP subtasks. This thesis work has made contributions to three main areas: sentiment lexicons 1.4.1, analysis of social media texts 1.4.2, and polarity classification 1.4.3. The last contribution of this thesis is to have put together a social media monitor based on the research done in the previous areas. We provide further details on the following subsections.

1.4.1 Polarity lexicons

We have analysed three strategies for less resourced languages. The initial approach was to translate existing lexicons by means of bilingual dictionaries. Results show that a great manual effort is needed to create accurate lexicons, due to translations not carrying the original polarity and many translations being infrequent words.

A second method was then researched by extracting polar words from corpora, following a strategy similar to (Turney and Littman, 2003). Even if the generated resources are noisy, since the cost of cleaning the lexicon is lower in this approach

and the extracted words tend to be frequent words, the resulting lexicons are rather useful.

Lastly, a fully automatic approach was taken, proposing QWN-PPV (San Vicente et al., 2014), a novel LKB graph-based method, by expanding the polarity of a few seeds by means of the UKB algorithm (Agirre and Soroa, 2009). Summarizing, it is specially important the pioneering work done for Basque. Specifically:

- Three sentiment lexicon creation methods for less resourced languages have been researched (see part I), proposing a novel fully automated approach (see chapter 3).
- The first Basque sentiment lexicons were generated by the work undertaken in this thesis (chapters 2 and 4).

1.4.2 Social media

This thesis addressed the two main challenges for processing social media data: language identification (see chapter 5) and microtext normalization (see chapter 6). The first one specially affects less resourced languages because they have little support in social media. Both tasks are approached in the same manner, namely by organizing a shared task to experience with problem at first hand, and comparing different techniques on controlled experimental environments. In addition, two different algorithms are proposed for microtext normalization, including one with resources for four languages (eu|es|en|fr) (chapters 7 and 10).

1.4.3 Polarity classification

Another important focus of this thesis has been multilingual polarity classification, again stressing the need to work on less resourced languages. Firstly, unsupervised polarity classifiers were implemented, mainly for the evaluation of the generated sentiment lexicons. With this aim in mind, we annotated small opinion test datasets for Basque.

Secondly, supervised classifiers were developed, with the objective of efficiently combining the information provided by a sentiment lexicon information with other linguistically motivated features. The proposed supervised classifiers were tested in several international shared tasks: the Spanish classifiers won twice the TASS shared task (Saralegi and San Vicente, 2012; Saralegi and San Vicente, 2013) and they

ranked second in the last participation (San Vicente and Saralegi, 2014); the English classifier obtained remarkable results in the Semeval 2015 aspect based sentiment analysis shared task (San Vicente et al., 2015).

Also, two new opinion annotated multilingual domain datasets have been compiled from Twitter, leading to the creation of the first supervised polarity classifiers for Basque (San Vicente et al., 2019).

The result of our research effort is a Multilingual SA system capable of analysing texts in four languages: Basque, English, French and Spanish (see part III).

1.4.4 Real World Application

Applying our research in a real scenario is also an important contribution. A fully open-sourced solution has been developed and successfully applied in two real scenarios: monitoring cultural events and political campaigns. This work can be seen as an example of technological transfer from the initial academic research stage to a final product development.

A crawler has been developed which serves two purposes: as a means to feed data into our social media monitor and as a tool to generate datasets from social media (chapter 10).

The software and resources developed during this thesis will be made public, if they are not already. Section 11.3 provides a detailed list of the software and resources resulting from this research.

1.5 Organization of the document

This dissertation presents chronologically the research we have carried out on Multilingual Sentiment Analysis in Social Media. From here on, the document is divided in five main parts.

The first three correspond to the main goals defined in section 1.3, starting from the first experiments on the construction of subjective vocabulary lexicons to a full multilingual polarity classification system, including language normalization modules for applying it to a social media environment. The 4th part addresses the application of the previous research in a real use case.

Parts I-IV are divided in chapters where each chapter corresponds to a published research paper. The last part deals with the conclusions of this thesis. Thus, the rest of this document is organized as follows:

- **Part I: Construction of Sentiment Lexicons**

  This part presents the experimental path we took in order to create sentiment lexicons we would later use to develop SA systems. All the experiments presented follow the premises that they should require minimum external resources and they should be cost-effective, making them suitable, not only for major languages but also for less resourced ones. This part is divided in three chapters: the first two (Saralegi et al., 2013; San Vicente et al., 2014) present various approaches to sentiment lexicon building, and the third chapter (San Vicente and Saralegi, 2016) presents a comparison between those approaches.

- **Part II: Social Media Analysis**

  The second part presents the challenges that must be faced when analysing social media data. The part is divided in three different chapters covering two main topics: language identification (Zubiaga et al., 2016) and language normalization of tweets (Saralegi and San Vicente, 2013b; Alegria et al., 2015).

- **Part III: Polarity Classification**

  Part III covers the research done on polarity classification. Two chapters present the experiments done for developing Spanish (San Vicente and Saralegi, 2014) and English (San Vicente et al., 2015) polarity classifiers, respectively. Those experiments are the foundation for the rest of the classifiers developed during this thesis.

- **Part IV: Real World Application**

  This last part presents all the previously acquired knowledge applied to a real use case scenario. Chapter 10 (San Vicente et al., 2019) describes in detail a real time social media monitor, a platform that allows automatic analysis of the impact in social media and digital press and of topics or domains specified by the user, based on NLP. Polarity lexicons, polarity annotated tweet datasets, polarity classification models and tweet normalization resources were developed in four languages: Basque, English, French and Spanish.

- **Part V: Conclusion and further work**

  Finally, we draw the main conclusions of this research and present some ideas for future work.

1.6 How to read this thesis

In the following, we provide a short index (section 1.6.1) pointing out to the various topics addressed in this thesis. The aim is to offer readers a quick guide on the different ways in which this thesis could be read.

This is complementary to table 1.1 which provides a summary of the most important contributions made by this thesis in a single page, and how to find them in the document.

1.6.1 Topic Index

- **Basque SA resources**.
  - Sentiment lexicons: chapter for subjectivity lexicons 2 and chapter 4 for polarity lexicons.
  - Basque polarity classifiers: chapter 2 for unsupervised classifiers and 10 for supervised classifiers.
  - Basque SA datasets: chapters 2 and 10.
  - Microtext normalization resources: chapter 10.
- **Less-resourced languages**.
  - Sentiment Lexicon construction:
    - Lexicon translation: chapters 2 and 4.
    - Polarity propagation from LKB: chapter 3 and 4.
  - Language identification: chapter 5.
  - Polarity classification and resources: chapter 10.
- **Microtext Normalization.** Chapters 6 and 7.
- **Shared tasks.**
  - Polarity Classification:
    - TASS (es): chapter 8.
    - SemEval ABSA (en): chapter 9.
  - Microtext normalization: TweetNorm, chapters 6 (task overview) and 7 (participation).
  - Language identification. TweetLID, chapter 5.
- **Sentiment lexicon evaluation**. Chapters 3 and 4.
- **Released resources and software**: Chapter 11, section 11.3.

1.6.2 Contribution Table

Part	Paper/Chapter	Topic(s)	Langs	Task	Datasets	Resources	Software
I	(Saralegi et al., 2013) Chapter 2	Subjectivity Lexicons - Translation, Corpus based	Eu	-	News, blogs, tweets, Music/Film reviews	Lexicons (eu, corpus based and translated)	DSPL
I	(San Vicente et al., 2014) Chapter 3	Sentiment Lexicons - LKB based	En, Es	-	-(Bespalov et al., 2011)* -MPQA* -HOpinion[17]*	-Lexicons (es,en) -MSOL*, General Inquirer*, SO-CAL*, Opinion Finder*, SentiWordnet*	QWN-PPV
I	(San Vicente and Saralegi, 2016) Chapter 4	Sentiment Lexicons - comparison	Eu	-	News, Music/Film reviews	-$ElhPolar_{eu}$ lexicon -QWN-PPV lexicons for Basque	-
II	(Zubiaga et al., 2016) Chapter 5	Language identification in Twitter	Ca, Gl, En, Es, Eu, Pt	TweetLID	TweetLID corpus	-	-
II	(Alegria et al., 2015) Chapter 6	Microtext Normalization	Es	TweetNorm	TweetNorm corpus	-	-
II	(Saralegi and San Vicente, 2013b) Chapter 7	Microtext Normalization	Es	TweetNorm	TweetNorm corpus*	OOV normalization dictionary (es, corpus-based)	Normalization module: heuristics + language models
III	(San Vicente and Saralegi, 2014) Chapter 8	Polarity classification	Es	TASS	TASS general*	$ElhPolar_{es}$ lexicon	SVM classifier
III	(San Vicente et al., 2015) Chapter 9	Polarity classification, Aspect Based SA	En	SemEval ABSA	SemEval ABSA 2015*	Sentiment Lexicons (en, domain specific)	EliXa
IV	(San Vicente et al., 2019) Chapter 10	Social Media monitor, normalization, Polarity classification	En, Es, Eu, Fr	-	-DSS2016 Behagunea -BEC2016 (politics)	Social media normalization resources	-Behagunea UI -MSM crawler -EliXa

Table 1.1: Organization of the document by contribution. Datasets or resources marked with an asterisk (*) were not developed during this thesis but they were used for evaluation purposes.

[17] `http://clic.ub.edu/corpus/hopinion`

PART I

CONSTRUCTION OF SENTIMENT LEXICONS

I Construction of Sentiment lexicons

One of the main resources of Sentiment Analysis are sentiment or polarity lexicons, namely, lists of words or lemmas annotated with prior polarities. Both supervised and unsupervised approaches have benefited from such resources, either to directly tag opinionated words (Taboada et al., 2011a; Thelwall, 2017), or to be used as features in machine learning approaches (Kouloumpis et al., 2011; Mohammad et al., 2013; Kiritchenko et al., 2014). Much research has been done on automatic methods to create such lexicons in order to avoid the high cost of manually created lexicons (Turney, 2002; Kaji and Kitsuregawa, 2006; Mihalcea et al., 2007; Pérez-Rosas et al., 2012; Esuli and Sebastiani, 2006). Then again, while automatic methods are cheaper, they often produce rather noisy resources.

All the experiments presented in this part follow the principle that they should require minimum external resources and that they should be cost-effective, making them suitable not only for major languages, but also for less resourced ones, such as Basque, one of the main aspects of our study. In the case of Basque, we started from scratch, building first subjectivity lexicons (limited to detect whether a piece of text is opinionated or not) and then moving forward towards polarity lexicons.

Chapter 2 (Saralegi et al., 2013) analyses two strategies for building subjectivity lexicons in an automatic way: translating existing subjectivity lexicons from a major language (English) into Basque, and building subjectivity lexicons from corpora. Subjective vocabulary was automatically inferred from various corpora. The lexicons are evaluated extrinsically in a subjectivity classification task, and intrinsically by manually reviewing the lexicons. Test datasets were annotated for several domains, at document- and/or sentence-level. Our experiments concluded that both methods may be adequate depending on the circumstances. Corpus-based lexicons perform better overall, although experiments showed that their performance varies significantly across domains.

The work carried out in (Saralegi et al., 2013) was the basis to generate polarity lexicons for Basque and Spanish. $ElhPolar_{es}$ (Saralegi and San Vicente, 2013) was generated by combining the two aforementioned approaches after adapting them to work with binary polarities (positive‖negative) and manually annotating the lexicon entry candidates. $ElhPolar_{eu}$ (San Vicente and Saralegi, 2015; San Vicente and Saralegi, 2016) was similarly created and it is used as part of EliXa.

The aforementioned strategies require either annotated corpora or manual effort to a certain extent. Thus, Chapter 3 (San Vicente et al., 2014) presents our approach to develop a method without those requirements. The result of our work was *QWN-PPV* (Q-WordNet as Personalized PageRanking Vector), a dictionary-based method requiring only a Lexical Knowledge Base (LKB). The generated lexicons outperform other automatically generated lexicons for various extrinsic evaluations. They are also competitive with respect to manually annotated ones. Results suggest that no single lexicon is best for every task and dataset and that the intrinsic evaluation of polarity lexicons is not a good indicator of their quality. The *QWN-PPV* method allows to easily create quality polarity lexicons whenever no domain-based annotated corpora are available for a given language.

With the previous approaches, we achieved our goal of providing methodologies to generate sentiment lexicons for less resourced languages. Still, our experience showed that we still spent time improving those lexicons manually before being used for polarity classification. Thus, a question remained unanswered: is the manual annotation worth, or are we spending too much effort to marginally improve the performance?

Chapter 4 (San Vicente and Saralegi, 2016) aims to answer this question by comparing the performance of three Basque sentiment lexicons in a polarity classification task : a) manually annotated lexicon; b) a corpus-based lexicons manually reviewed and, c) a completely automatic lexicon generated by the QWN-PPV method. Results show that the corpus-based method is the best option if we can afford some human effort, but completely automatic methods should not be disregarded, as they are the most cost efficient.

The list of the publications on this part of the thesis by order of appearance if provided below. Furthermore, with each publication we provide details regarding the contribution of the author of this thesis.

- Xabier Saralegi, Iñaki San Vicente, and Irati Ugarteburu. Cross-lingual projections vs. corpora extracted subjectivity lexicons for less-resourced languages. In Alexander Gelbukh, editor, *Computational Linguistics and*

Intelligent Text Processing, volume 7817 of *Lecture Notes in Computer Science*, pages 96–108. 2013. ISBN 978-3-642-37255-1

Contribution to the paper: First and second author contributed equally to the paper. Iñaki San Vicente was responsible for the creation of the various lexicons analysed, and the source code for both for lexicon generation and also and dataset evaluation. Xabier Saralegi contributed to the design of the experiments, analysis of the results and writing of the paper. Irati Ugarteburu helped with the manual annotation of subjectivity lexicons. All three authors took part in the annotation of the evaluation datasets.

- Iñaki San Vicente, Rodrigo Agerri, and German Rigau. Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 88–97, 2014

  Contribution to the paper: Main author of the paper. Responsible for all the coding and data processing. Second and third authors contributed to the design of the experiments, result analysis and the writing of the paper.

- Iñaki San Vicente and Xabier Saralegi. Polarity lexicon building: to what extent is the manual effort worth? In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may 2016. ISBN 978-2-9517408-9-1

  Contribution to the paper: Main author of the paper. Responsible for the coding and data processing. Both authors contributed equally in the design of the experiments, as well as in the writing of the paper. Manual annotation of the lexicons was produced by both authors.

CHAPTER 2

Corpus-based lexicons

Cross-Lingual Projections vs. Corpora Extracted Subjectivity Lexicons for Less-Resourced Languages

Xabier Saralegi, Iñaki San Vicente, Irati Ugarteburu

Elhuyar Foundation

Subjectivity tagging is a prior step for sentiment annotation. Both machine learning based approaches and linguistic knowledge based ones profit from using subjectivity lexicons. However, most of these kinds of resources are often available only for English or other major languages. This work analyzes two strategies for building subjectivity lexicons in an automatic way: by projecting existing subjectivity lexicons from English to a new language, and building subjectivity lexicons from corpora. We evaluate which of the strategies performs best for the task of building a subjectivity lexicon for a less resourced language (Basque). The lexicons are evaluated in an extrinsic manner by classifying subjective and objective text units belonging to various domains, at document- or sentence-level. A manual intrinsic evaluation is also provided which consists of evaluating the correctness of the words included in the created lexicons.

Published in *Computational Linguistics and Intelligent Text Processing, volume 7817 of Lecture Notes in Computer Science*, pages 96–108. 2013. ISBN 978-3-642-37255-1.

2.1 Introduction

Opinion mining or sentiment analysis are tasks involving subjectivity detection and polarity estimation. Both tasks are necessary in many sentiment analysis applications, including sentiment aggregation and summarization or product comparisons. Researchers into sentiment analysis have pointed out the frequent benefit of a two-stage approach, in which subjective instances are distinguished from objective ones, after which the subjective instances are further classified according to polarity (Yu and Hatzivassiloglou, 2003; Pang and Lee, 2004; Wilson et al., 2005). Pang and Lee (2004) obtain an improvement from 82.8% to 86.4% for polarity classification by applying a subjectivity classifier in advance. So, developing a method for subjectivity detection seems an adequate first step for building an Opinion mining system for a certain language.

When dealing with subjectivity, some authors proposed rule-based methods (Riloff and Wiebe, 2003b) which use subjectivity lexicons. Other authors propose supervised methods based on machine learning techniques (Yu and Hatzivassiloglou, 2003). In both cases, subjectivity lexicons are an important knowledge resource. So it is clear that subjectivity lexicons are a key resource for tackling this task. Nowadays, there are widely used lexicons, such as OpinionFinder (Wiebe et al., 2005), Sentiwordnet (Esuli and Sebastiani, 2006) and General Inquirer (Stone et al., 1966), but, as is the case with many NLP resources, those lexicons are geared towards major languages. This means that new subjectivity lexicons must be developed when dealing with many other languages.

As manual building is very costly and often uneconomic for most languages, especially less resourced languages, machine building methods offer a viable alternative. In that sense, several methods (Turney, 2002; Kaji and Kitsuregawa, 2007; Mihalcea et al., 2007; Banea et al., 2008; Wan, 2008) have been proposed for building subjectivity lexicons. The methods rely on two main strategies: building the lexicon from corpora or trying to project existing subjectivity resources to a new language. The first approach often produces domain specific results, and so, its performance in out-of-domain environments is expected to be poorer. Projecting a lexicon to another language would produce a resource that would *a priori* be more

consistent in all environments. However, as the projection involves a translation process, the errors ocurring at that step could reduce the quality of the final lexicon as shown by Mihalcea et al. (2007).

In our research we compared these two cost-effective strategies for building a subjectivity lexicon for a less resourced language. We assumed that for languages of this type the availability of parallel corpora and MT systems is very limited, and that was why we avoided using such resources. Our contribution lies in a robust cross-domain evaluation of the two strategies. This experiment was carried out using Basque. First, we compared the correctness of the resulting lexicons at word level. Then, the lexicons were applied in a task to classify subjectivity and objectivity text units belonging to different domains: newspapers, blogs, reviews, tweets and subtitles.

The paper is organized as follows. The next chapter offers a brief review of the literature related to this research, and discusses the specific contributions of this work. The third section presents the resources we used for building the subjectivity lexicons, the experiments we designed and the methodology we followed. In the fourth chapter, we describe the different evaluations we carried out and the results obtained. Finally, some conclusions are drawn and we indicate some future research directions.

2.2 State of the Art

Wilson et al. (2005) define a subjective expression as any word or phrase used to express an opinion, emotion, evaluation, stance, speculation, etc. A general covering term for such states is private state. Quirk et al. (1985) define a private state as a state that is not open to objective observation or verification: “a person may be observed to assert that God exists, but not to believe that God exists”. Belief is in this sense ’private’. So, subjectivity tagging or detection consists of distinguishing text units (words, phrases sentences...) used to present opinions and other forms of subjectivity from text units used to objectively present factual information. Detection is part of a more complex task which Wilson (2008) called subjectivity analysis, which consists of determining when a private state is being expressed and identifying the attributes of that private state. Identifying attributes such as the target of the opinion, the polarity of the subjective unit or its intensity, is outside the range of this work.

2.2.1 Subjectivity detection methods

Methods for subjectivity detection can be divided into two main approaches. Rule-based methods which rely on subjectivity lexicons, and supervised methods based on classifiers trained from annotated corpora.

Wiebe et al. (1999) use manually annotated sentences for training Naive Bayes classifiers. Pang and Lee (2004) successfully apply Naive Bayes and SVMs for classifying sentences in movie reviews. Wang and Fu (2010) present a sentiment density-based naive Bayesian classifier for Chinese subjectivity classification. Das and Bandyopadhyay (2009b) propose a Conditional Random Field (CRF)-based subjectivity detection approach tested on English and Bengali corpora belonging to multiple domains.

Lexicon-based systems are also proposed in the literature. Turney (2002) computed the average semantic orientation of product reviews based on the orientation of phrases containing adjectives and adverbs. The classifier proposed by Riloff and Wiebe (2003b) uses lists of lexical items that are good subjectivity clues. It classifies a sentence as subjective if it contains two or more of the strongly subjective clues. Das and Bandyopadhyay (2009a) proposed a classifier which uses sentiment lexicons, theme clusters and POS tag labels.

A third alternative would be to combine both approaches. Yu and Hatzivassiloglou (2003) obtain 97% precision and recall using a Bayesian classifier that uses lexical information. This proves that subjectivity lexicons are indeed important resources.

According to Yu and Kübler (2011), opinion detection strategies designed for one data domain generally do not perform well in another domain, due to the variation of the lexicons across domains and different registers. They evaluated the subjectivity classification in news articles, semi-structured movie reviews and blog posts using Semi-Supervised Learning (SSL) methods, and obtained results that vary from domain to domain. Jijkoun and de Rijke (2011) propose a method to automatically generate subjectivity clues for a specific topic by extending a general purpose subjectivity lexicon.

2.2.2 Methods for subjectivity lexicon building

Text corpora are useful for obtaining subjectivity and polarity information associated with words and phrases. Riloff et al. (2003) adopt a bootstrapping strategy based on patterns to extend a seed set of 20 terms classified as strongly subjective. Baroni and Vegnaduzzo (2004) apply the PMI (Pointwise Mutual Information) method

to determine term subjectivity. Subjectivity level is measured according to the association degree with respect to a seed set of 35 adjectives marked as subjective.

When tackling the problem of the lack of annotated corpora, many authors propose using MT techniques. Mihalcea et al. (2007) annotate an English corpus using OpinionFinder Wiebe et al. (2005) and use cross-lingual projection across parallel corpora to obtain a Romanian corpus annotated for subjectivity. Following the same idea, Banea et al. (2008) use machine translation to obtain the required parallel corpora. In this case they apply the method for Romanian and Spanish. Wan (2008) also proposed the generation of Chinese reviews from English texts by Machine Translation.

Another approach to building a subjective word list in a language is the translation of an existing source language lexicon by using a bilingual dictionary. Mihalcea et al. (2007) used a direct translation process to obtain a subjectivity lexicon in Romanian. Their experiments concluded that the Romanian subjectivity clues derived through translation are less reliable than the original set of English clues, due to ambiguity errors in the translation process. Das and Bandyopadhyay (2009b) proposed improving the translation of ambiguous words by using a stemming cluster technique followed by SentiWordNet validation. Jijkoun and Hofmann (2009) apply a PageRank-like algorithm to expand the set of words obtained through machine translation.

Banea et al. (2011) compare different methods of subjectivity classification for Romanian. Among subjectivity lexicon building methods, there are bootstrapping a lexicon by using corpus-based word similarity, and translating an existing lexicon. They conclude that the corpus-based bootstrapping approach provides better lexicons than projection.

In this work we wanted to analyze strategies for developing a subjectivity lexicon for a Less-Resourced Language. We assumed that such languages can only avail themselves of monolingual corpora and bilingual lexicons. So parallel corpora, MT system-based approaches and approaches based on large subjectivity annotated corpora are not contemplated. We focused on a corpus-based approach and projection onto the target language.

2.3 Experiments

Projection-based lexicon building requires a subjectivity lexicon L_S_s in a source language s and a bilingual dictionary $D_{s \to t}$ from s to the target language t. In our

experiments we took the English subjectivity lexicon (L_S_{en}) introduced in Wiebe et al. (2005) as a starting point. L_S_{en} contains 6,831 words (4,743 strong subjective and 2,188 weak subjective). According to the authors, those subjective words were collected from manually developed resources and also from corpora. Strong subjective clues have subjective meanings with high probability, and weak subjective clues have a lower probability of having subjective meanings. As for the bilingual dictionary, a bilingual English-Basque dictionary $D_{en \rightarrow eu}$ which includes 53,435 pairs and 17,146 headwords was used.

Corpora-based lexicon extraction requires subjective and objective corpora. Subjective and objective corpora can be built by using simple heuristics. News from newspapers or Wikipedia articles can be taken as objective documents. Opinion articles from newspapers can be taken as subjective articles. Those heuristics are not trouble free, but then again, they allow us to create low-cost annotated corpora. Using news as an objective corpus can be a rough heuristic because, according to Wiebe et al. (2001), many sentences (44%) included in news are subjective. On the other hand, as Wikipedia belongs to a different domain from that of newspaper opinion articles, some divergent words can be incorrectly identified as subjective if we compare a Wikipedia corpus with a subjective corpus comprising opinion articles, due to the fact that they are a feature in the journalism domain but not in Wikipedia texts.

We built a subjective corpus TC_S_{eu} by taking 10,661 opinion articles from the Basque newspaper Berria[1]. Two objective corpora were built: one by collecting 50,054 news items from the same newspaper TCN_O_{eu}, and the other by gathering all the articles (143,740) from the Basque Wikipedia TCW_O_{eu}. A subset of TCN_O_{eu} containing the same number of articles as TC_S_{eu} was also prepared for parameter tuning purposes which we will name $TCN\_O'_{eu}$.

2.3.1 Cross-lingual projection of the subjectivity lexicon

We translated the English subjectivity lexicon L_S_{en} by means of a bilingual dictionary $D_{en \rightarrow eu}$ to create a Basque subjectivity lexicon L_P_{eu}. Ambiguities are resolved by taking the first translation[2]. Using this method we obtained translations for 36.67% of the subjective English words: L_P_{eu} includes 1,402 strong and 1,169 weak subjective words. The number of translations obtained was low, especially

[1] http://berria.info

[2] The bilingual dictionary has its translations sorted according to their frequency of use, so the first translation method should provide us with the most common translations of the source words.

for strong subjective words. Most of these words are inflected (e.g., *"terrified"*, *"winners"*, ...) forms or derived words where prefixes or suffixes have been added (e.g., *"inexact"*, *"afloat"*, ...).

According to Mihalcea et al. (2007) translation ambiguity is another problem that distorts the projection process. In their experiments Romanian subjectivity clues derived through translation were less reliable than the original set of English clues. In order to measure to what extent that problem would affect our projection, we randomly selected 100 English words and their corresponding translations. Most of the translations (93%) were correct and subjective according to a manual annotation involving two annotators (97% inter-tagger agreement, Cohen's k=0.83). So we can say that the translation selection process is not critical. We annotated as correct translations those corresponding to the subjective sense of the English source word. Unlike Mihalcea et al. (2007), we did not analyze whether the translated word had less subjective connotation than the source word.

2.3.2 Corpus-based lexicon building

Our approach was based on inferring subjective words from a corpus which includes subjective and objective documents. So, we identified as subjective words those whose relevance in subjective documents is significantly higher than in objective documents. We adopted a corpus-based strategy, because it is affordable and easily applicable to less resourced languages. We extracted Basque subjectivity lexicons in accordance with various relevance measures and objective corpora. TC_S_{eu} was used as the subjective corpus, and TCW_O_{eu} (Wikipedia) or TCN_O_{eu} (News) as objective corpora. For each word w in the subjective corpus we measured its degree of relevance with respect to the subjective corpus as compared with the objective corpus. That way we obtained the most salient words in a certain corpus, the subjective corpus in this case. We took that degree of relevance as the subjectivity degree $bal(w)$. That degree was calculated by the Log Likelihood ratio (LLR) or by the percentage difference ($\%DIFF$). Maks and Vossen (2012) compared LLR and $\%DIFF$ for that purpose, and obtained better results by using $\%DIFF$.

In order to evaluate the adequacy of the measurements (LLR or $\%DIFF$) and the various corpus combinations (Wikipedia or News for the objective part), we analyzed how subjective and objective words are distributed through the rankings corresponding to the different combinations (*LLR_News*, *DIFF_News*, *DIFF_Wiki* and *LLR_Wiki*). For that aim, two references were prepared. The first one includes only subjective words, while the second one includes both objective and subjective

words. The first reference was built automatically by taking the strong subjective words of L_P_{eu}. For the second reference three annotators manually tagged subjective and objective words in a sample of 500 words selected randomly from the intersection of all candidate dictionaries (*DIFF_Wiki*, *DIFF_News*, *LLR_Wiki* and *LLR_News*). The overall inter-agreement between the annotators was 81.6% (Fleiss' k=0.63). Simple majority was used for resolving disagreements (27% of the words evaluated).

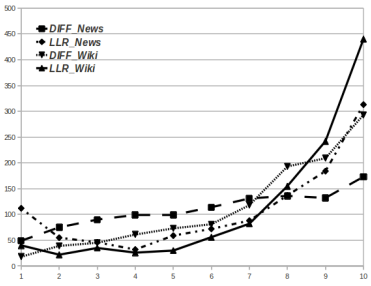


Figure 2.1: Distribution of subjective words with various measures and corpus combinations

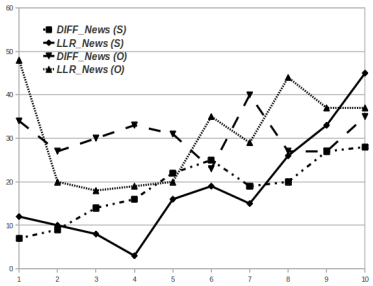


Figure 2.2: Distribution of subjective and objective words using TCN_O_{eu} as objective corpus.

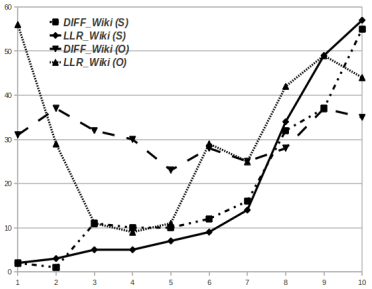


Figure 2.3: Distribution of subjective and objective words using TCW_O_{eu} as objective corpus.

According to the results shown in Figures 2.1, 2.2 and 2.3 Wikipedia seems to be a more adequate objective corpus. It provides a higher concentration of subjective words in the first positions of the rankings[3] (i.e. last intervals) than News when using both measurements and for both references. In addition, the concentration of objective words in the first positions is slightly lower when using TCW_O_{eu}, compared with using TCN_O_{eu} as the objective reference corpus.

Regarding the measurements, LLR provides better distributions of subjective words than $\%DIFF$ for both reference corpora. The highest concentration of the subjective words is in the first positions of the rankings. However $\%DIFF$ seems to be more efficient for removing objective ones from first ranking positions. Figure 2.4 plots the distribution of subjective/objective word rates across different ranking intervals. The best ratio distribution is achieved by the $\%DIFF$ measurement when used in combination with TCW_O_{eu}.

In terms of size, corpora-based lexicons are bigger than the projection-based one. For high confidence thresholds, $LLR > 3.84$, $p\text{-}value < 0.05$; and $\%DIFF > 100$ (Maks and Vossen, 2012), corpora-based lexicons provide 9,761; 6,532; 8,346 and

[3]In Figures 2.1, 2.2, 2.3 and 2.4, higher intervals contain words scoring higher in the rankings.

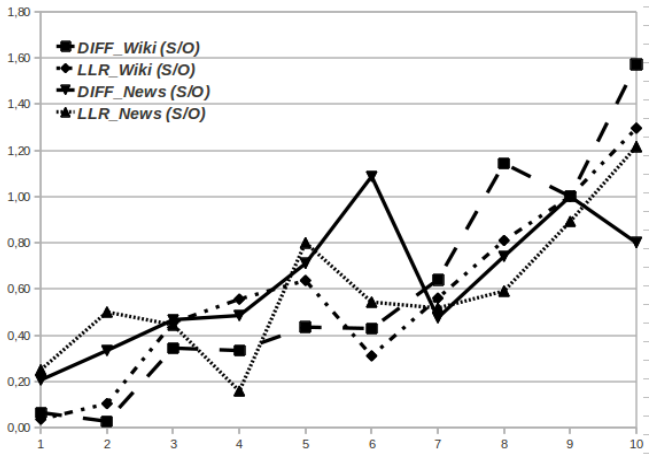


Figure 2.4: Subjective/objective ratio with respect to ranking intervals.

6,748 words for *DIFF_Wiki*, *DIFF_News*, *LLR_Wiki* and *LLR_News*, respectively. These will be the dictionaries used in the evaluation presented in the next section. The sizes of these dictionaries are close to that of the source English lexicon L_S_{en} (6,831 words). However, after projecting it to Basque, this number goes down to 2,571. So it seems that the corpora-based strategy provides bigger subjectivity lexicons. Then again, we have to take into account that corpus-based lexicons include several objective words (See Figure 1.). In addition, corpus-based lexicons are biased towards the domain of journalism.

2.4 Evaluation

2.4.1 Classifier

In this work, we adopted a simple lexicon-based classifier similar to the one proposed in (Wang and Liu, 2011). We propose the following ratio for measuring the subjectivity of a text unit tu:

$$subrat(tu) = \sum_{w \in tu} bal(w)/|tu| \tag{2.1}$$

where $bal(w)$ is 1 if w is included in the subjectivity lexicon[4].

Those units that reach a threshold are classified as subjective. Otherwise, the units are taken as objective. Thresholds are tuned by maximising accuracy when classifying the training data at document level. Even if most of the evaluation data collections are tagged at sentence level, the lack of a sentence level annotated training corpus led us to choose this parameter optimisation method. In order to tune the threshold with respect to a balanced accuracy for subjective and objective classification, tuning is done with respect to a balanced training corpus comprising TC_S_{eu} and $TCN\_O'_{eu}$, which we will call $Train\_D$.

2.4.2 Annotation Scheme

We evaluated the subjectivity lexicons obtained by the different methods in an extrinsic manner by applying them within the framework of a classification task. That way we measured the adequacy of each lexicon in a real task. The gold-standard used for measuring the performance comprises subjective and objective text units that belong to different domains. As we mentioned in section 2.2.1, the performance of subjectivity classification systems is very sensitive to the application domain. In order to analyze that aspect, we prepared the following test collections:

- Journalism documents ($Jour\_D$) and sentences ($Jour\_S$): texts collected from the Basque newspaper Gara[5].
- Blog sentences ($Blog\_S$): texts collected from Basque blogs included in the website of Berria.
- Twitter sentences ($Tweet\_S$): tweets collected from the aggregator of Basque tweets Umap[6]. Only tweets written in standard Basque are accepted.
- Sentences of music reviews (Rev_S): reviews collected from the Gaztezulo[7] review site.

[4]We experimented using weights based on the strength of subjectivity but no improvement was achieved, and so, these results are not reported.

[5]http://www.gara.net

[6]http://umap.eu/

[7]http://www.gaztezulo.com/

- Sentences of subtitles (*Sub_S*): subtitles of different films are collected from the azpitituluak.com site.

In the case of documents, no manual annotation was done. Following the method explained in section 2.3, we regarded all opinion articles as subjective, and all news articles as objective. The sentences were manually annotated. Our annotation scheme is simple compared to that used in MPQA (Wiebe et al., 2005) which represents private states and attributions. In contrast, our annotation is limited to tagging a sentence as subjective if it contains one or more private state expression; otherwise, the sentence is objective. A private state covers opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgements.

Source	Unit	Domain	#units	#sub+	#sub	#obj	#obj+
Train_D	document	Journalism	21,320	10,660		10,660	
Jour_D	document	Journalism	9,338	4,669		4,669	
Jour_S	sentence	Journalism	192	60	46	35	51
Blog_S	sentence	Blog	206	94	50	20	42
Tweet_S	sentence	Twitter	200	69	40	21	70
Rev_S	sentence	Music Reviews	138	54	36	24	24
Sub_S	sentence	Subtitles	200	98	31	20	51

Table 2.1: Statistics and class distribution of the reference collections.

We classified sentences according to four categories, depending on aspects such as the number of private state expressions, their intensity, etc.: completely subjective (sub+); subjective but containing some objective element (sub); mostly objective but containing some subjective element (obj); and completely objective (obj+). In order to obtain a robust annotation, three references per annotation were done by three different annotators. Disagreement cases were solved in two different ways. Firstly, annotators discussed all sentences including three different annotations or two equal annotations and a third that was to a distance of more than one category, until consensus was achieved. For dealing with the rest of the disagreement cases, majority voting was used. Table 2.1 shows the statistics for the test collections and the results of our annotation work.

2.4.3 Results

By means of our average ratio classifier, we classified the text units in the seven collections presented in the previous section. As mentioned in section 2.4.1, the units in the test collections were classified according to the subjectivity threshold tuned over the documents in *Train_D*. The optimum subjectivity threshold is computed for each lexicon we evaluated (L_P_{eu}, *DIFF_News*, *LLR_News*, *DIFF_Wiki* and *LLR_Wiki*).

	L_P_{eu}	*DIFF_Wiki*	*DIFF_News*	*LLR_Wiki*	*LLR_News*
Train_D	0.63	0.66	**0.90**	0.64	0.87
Jour_D	0.74	0.76	0.80	0.74	**0.87**
Jour_S	0.63	0.59	0.57	0.58	**0.64**
Blog_S	0.65	**0.73**	0.66	**0.73**	0.72
Tweet_S	**0.68**	0.58	0.62	0.59	0.60
Rev_S	**0.70**	**0.70**	0.67	0.67	0.67
Sub_S	0.67	**0.71**	0.70	0.67	0.67

Table 2.2: Accuracy results for subjectivity and objectivity classification.

Table 2.2 and 2.3 present overall accuracy results and F-score results of the subjective units achieved by the different lexicons in the various test collections. In this evaluation, only a binary classification was performed, text units belonging to **obj** and **obj+** classes were grouped into a single category, and the same was done for **sub** and **sub+**. Firstly, according to those results, corpus-based lexicons compiled using TCN_O_{eu} (News) as objective reference (columns 3 and 5) are very effective for document classification. The projected lexicon L_P_{eu} performs significantly worse. Those results were expected, since the corpora-based lexicons have the domain advantage. However, L_P_{eu}'s performance is comparable to corpus-based lexicons' on non-journalistic domains. Moreover, it is better than the corpus-based lexicons in the Twitter domain, both in terms of accuracy and F-score of subjective units. Taking all the results into account, we can see that despite the better performance of corpus-based lexicons in most the domains, the performance of the projected lexicon is more stable across domains than the performance of corpus-based lexicons.

	L_P_{eu}	*DIFF_Wiki*	*DIFF_News*	*LLR_Wiki*	*LLR_News*
$Train\_D$	0.65	0.68	**0.90**	0.68	0.87
$Jour\_D$	0.75	0.77	0.82	0.75	**0.86**
$Jour\_S$	0.73	0.71	0.58	0.72	**0.74**
$Blog\_S$	0.76	0.82	0.77	**0.83**	**0.83**
$Tweet\_S$	**0.73**	0.69	0.70	0.70	0.71
Rev_S	0.79	0.77	0.78	0.75	**0.80**
Sub_S	0.78	**0.81**	0.79	0.78	0.79

Table 2.3: F-score results for subjectivity classification.

With regard to the corpus used as objective reference (columns 2 and 4 versus columns 3 and 5), the use of the wikipedia corpus TCW_O_{eu} improves the results of the News corpus only in non-journalistic domains and in terms of accuracy. Furthermore, Table 2.3 shows that if we only take into account the classification of subjective text units, TCN_O_{eu} performs better in all cases except for the subtitle domain collection.

Differences between LLR and $\%DIFF$ vary across the domains. In terms of accuracy, $\%DIFF$ provides better performance when dealing with tweets, reviews, and subtitles. On the contrary, in terms of F-score of subjective units, $\%DIFF$ is only better over subtitles.

We used 4 categories to annotate the references with different degrees of subjectivity. It is interesting how the performance of subjectivity detection changes depending on the required subjectivity degree. In some scenarios only the detection of highly subjective expressions is demanded. In order to adapt the system to those scenarios, we optimised the subjectivity threshold by maximising the $F_{0.5}$-score against training data. Table 2.4 shows precision and recall results for subjectivity detection if we only accept the ones that belong to the class **sub+** as subjective sentences. According to those results, with the new optimisation of the threshold, the system's performance for classifying **sub+** is similar to that of the initial system.

	L_P_{eu} sub+			LLR_News sub+		
	P	R	F	P	R	F
$Jour\_S$	0.61	0.90	0.73	0.65	0.84	0.73
$Blog\_S$	0.73	0.80	0.76	0.74	0.96	0.83
$Tweet\_S$	0.67	0.82	0.73	0.64	0.83	0.72
Rev_S	0.73	0.86	0.79	0.65	0.99	0.79
Sub_S	0.69	0.88	0.78	0.68	0.99	0.80

Table 2.4: Precision, recall and F-score results for detecting clearly subjective sentences.

2.5 Conclusions and future work

This paper has presented the comparison between two techniques to automatically build subjectivity lexicons. Both techniques only rely on easily obtainable resources, and are adequate for less resourced languages.

Our results show that subjectivity lexicons extracted from corpora provide a higher performance than the projected lexicon over most of the domains. Accuracies obtained with this method range from 87%, in case of the document classification, to 60-67%, in case of sentences. Projection provides a slight better performance only when dealing with non-journalistic domains. So, it could be an alternative for those domains. If we are interested in identifying only very subjective sentences, both methods offer a good performance (0.72-0.83 in terms of F-score), in particular, the corpora extracted subjectivity lexicons. Hence, the resources obtained with our methods could be applied in social-media analysis tasks where precision is the priority.

Regarding to ongoing and future work, as we have already mentioned, the methods we have researched in this paper are applicable to less resourced languages because they only require widely available resources. At the moment, we are analyzing the effect the characteristics (size, domain,...) of the resources used have on the quality of the final subjectivity lexicon. In the future, we plan to evaluate the Bootstrapping method proposed by Banea et al. 2008, which also relies on corpora.

Acknowledgements.

This work has been partially funded by the Industry Department of the Basque Government under grants IE12-333 (Ber2tek project) and SA-2012/00180 (BOM2 project).

CHAPTER 3

Dictionary-based Lexicons

Simple, Robust and (almost) Unsupervised Generation of Polarity Lexicons for Multiple Languages

Iñaki San Vicente, Rodrigo Agerri, German Rigau

IXA Group - University of the Basque Country (UPV/EHU)

This paper presents a simple, robust and (almost) unsupervised dictionary-based method, QWN-PPV *(Q-WordNet as Personalized PageRanking Vector) to automatically generate polarity lexicons. We show that* QWN-PPV *outperforms other automatically generated lexicons for the four extrinsic evaluations presented here. It also shows very competitive and robust results with respect to manually annotated ones. Results suggest that no single lexicon is best for every task and dataset and that the intrinsic evaluation of polarity lexicons is not a good performance indicator on a Sentiment Analysis task. The* QWN-PPV *method allows to easily create quality polarity lexicons whenever no domain-based annotated corpora are available for a given language.*

Published in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, April 26-30, Gothenburg, Sweden, pages 88–97, 2014.

3.1 Introduction

Opinion Mining and Sentiment Analysis are important for determining opinions about commercial products, on companies reputation management, brand monitoring, or to track attitudes by mining social media, etc. Given the explosion of information produced and shared via the Internet, it is not possible to keep up with the constant flow of new information by manual methods.

Sentiment Analysis often relies on the availability of words and phrases annotated according to the positive or negative connotations they convey. 'Beautiful', 'wonderful', and 'amazing' are examples of positive words whereas 'bad', 'awful', and 'poor' are examples of negatives.

The creation of lists of sentiment words has generally been performed by means of manual-, dictionary- and corpus-based methods. Manually collecting such lists of polarity annotated words is labor intensive and time consuming, and is thus usually combined with automated approaches as the final check to correct mistakes. However, there are well known lexicons which have been fully (Stone et al., 1966; Taboada et al., 2011b) or at least partially *manually created* (Hu and Liu, 2004b; Riloff and Wiebe, 2003a).

Dictionary-based methods rely on some dictionary or lexical knowledge base (LKB) such as WordNet (Fellbaum and Miller, 1998) that contain synonyms and antonyms for each word. A simple technique in this approach is to start with some sentiment words as seeds which are then used to perform some iterative propagation on the LKB (Hu and Liu, 2004b; Strapparava and Valitutti, 2004; Kim and Hovy, 2004a; Takamura et al., 2005; Turney and Littman, 2003; Mohammad et al., 2009; Agerri and García-Serrano, 2010; Baccianella et al., 2010).

Corpus-based methods have usually been applied to obtain domain-specific polarity lexicons: they have been created by either starting from a seed list of known words and trying to find other related words in a corpus or by attempting to directly adapt a given lexicon to a new one using a domain-specific corpus (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Ding et al., 2008; Choi and Cardie, 2009; Mihalcea et al., 2007). One particular issue arising from corpus methods is that for a given domain the same word can be positive in one context but negative in another. This is also a problem shared by manual and dictionary-based methods, and that is why *QWN-PPV* also produces synset-based lexicons for approaches on Sentiment Analysis at sense level.

This paper presents a simple, robust and (almost) unsupervised dictionary-based method, *QWordNet-PPV* (QWordNet by Personalized PageRank Vector) to

automatically generate polarity lexicons based on propagating some automatically created seeds using a Personalized PageRank algorithm (Agirre et al., 2014; Agirre and Soroa, 2009) over a LKB projected into a graph. We see *QWN-PPV* as an effective methodology to easily create polarity lexicons for any language for which a WordNet is available.

This paper empirically shows that: (i) *QWN-PPV* outperforms other automatically generated lexicons (e.g. SentiWordNet 3.0, MSOL) on the 4 extrinsic evaluations presented here; it also displays competitive and robust results also with respect to manually annotated lexicons; (ii) no single polarity lexicon is fit for every Sentiment Analysis task; depending on the text data and the task itself, one lexicon will perform better than others; (iii) if required, *QWN-PPV* efficently generates many lexicons on demand, depending on the task on which they will be used; (iv) intrinsic evaluation is not appropriate to judge whether a polarity lexicon is fit for a given Sentiment Analysis (SA) task because good correlation with respect to a *gold-standard* does not correspond with *correlation* with respect to a SA task; (v) it is easily applicable to create *qwn-ppv(s)* for *other languages*, and we demonstrate it here by creating many polarity lexicons not only for English but also for Spanish; (vi) the method works at *both word and sense* levels and it only requires the availability of a LKB or dictionary; finally, (vii) a dictionary-based method like *QWN-PPV* allows to easily create quality polarity lexicons whenever no domain-based annotated reviews are available for a given language. After all, there usually is available a dictionary for a given language; for example, the Open Multilingual WordNet site lists WordNets for up to 57 languages (Bond and Foster, 2013).

Although there has been previous work using graph methods for obtaining lexicons via propagation, the *QWN-PPV* method to combine the seed generation and the Personalized PageRank propagation is novel. Furthermore, it is considerable simple and obtains better and easier to reproduce results than previous automatic approaches (Esuli and Sebastiani, 2007; Mohammad et al., 2009; Rao and Ravichandran, 2009).

Next section reviews previous related work, taking special interest on those that are currently available for evaluation purposes. Section 3.3 describes the *QWN-PPV* method to automatically generate lexicons. The resulting lexical resources are evaluated in section 3.4. We finish with some concluding remarks and future work in section 3.5.

3.2 Related Work

There is a large amount of work on Sentiment Analysis and Opinion Mining, and good comprehensive overviews are already available (Pang and Lee, 2008; Liu, 2012), so we will review the most representative and closest to the present work. This means that we will not be reviewing corpus-based approaches but rather those constructed manually or upon a dictionary or LKB. We will in turn use the approaches here reviewed for comparison with *QWN-PPV* in section 3.4.

The most popular manually-built polarity lexicon is part of the General Inquirer (Stone et al., 1966), and consists of 1915 words labelled as "positive" and 2291 as "negative". Taboada et al. (2011b) manually created their lexicons annotating the polarity of 6232 words on a scale of 5 to -5. Liu *et al.*, starting with (Hu and Liu, 2004b), have along the years collected a manually corrected polarity lexicon which is formed by 4818 negative and 2041 positive words. Another manually corrected lexicon (Riloff and Wiebe, 2003a) is the one used by the Opinion Finder system (Wilson et al., 2005) and contains 4903 negatively and 2718 positively annotated words respectively.

Among the automatically built lexicons, (Turney and Littman, 2003) proposed a minimally supervised algorithm to calculate the polarity of a word depending on whether it co-ocurred more with a previously collected small set of positive words rather than with a set of negative ones. Agerri and García Serrano presented a very simple method to extract the polarity information starting from the *quality* synset in WordNet (Agerri and García-Serrano, 2010). Mohammad et al. (2009) developed a method in which they first identify (by means of affixes rules) a set of positive/negative words which act as seeds, then used a Roget-like thesaurus to mark the synonymous words for each polarity type and to generalize from the seeds. They produce several lexicons the best of which, MSOL(ASL and GI) contains 51K and 76K entries respectively and uses the full General Inquirer as seeds. They performed both intrinsic and extrinsic evaluations using the MPQA 1.1 corpus.

Finally, there are two approaches that are somewhat closer to us, because they are based on WordNet and graph-based methods. SentiWordNet 3.0 (Baccianella et al., 2010) is built in 4 steps: (i) they select the synsets of 14 paradigmatic positive and negative words used as seeds (Turney and Littman, 2003). These seeds are then iteratively extended following the construction of WordNet-Affect (Strapparava and Valitutti, 2004). (ii) They train 7 supervised classifiers with the synsets' glosses which are used to assign *polarity* and *objectivity* scores to WordNet senses. (iii) In SentiWordNet 3.0 (Esuli and Sebastiani, 2007) they take the output of the

supervised classifiers as input to applying PageRank to WordNet 3.0's graph. (iv) They intrinsically evaluate it with respect to MicroWnOp-3.0 using the *p-normalized Kendall τ distance* (Baccianella et al., 2010). Rao and Ravichandran (2009) apply different semi-supervised graph algorithms (Mincuts, Randomized Mincuts and Label Propagation) to a set of seeds constructed from the General Inquirer. They evaluate the generated lexicons intrinsically taking the General Inquirer as the gold standard for those words that had a match in the generated lexicons.

In this paper, we describe two methods to automatically generate seeds either by following Agerri and García-Serrano (2010) or using Turney and Littman's (2003) seeds. The automatically obtained seeds are then fed into a Personalized PageRank algorithm which is applied over a WordNet projected on a graph. This method is fully automatic, simple and unsupervised as it only relies on the availability of a LKB.

3.3 Generating qwn-ppv

The overall procedure of our approach consists of two steps: **(1)** automatically creates a set of seeds by iterating over a LKB (e.g. a WordNet) relations; and **(2)** uses the seeds to initialize contexts to propagate over the LKB graph using a Personalized Pagerank algorithm. The result is *qwn-ppv(s)*: Q-WordNets as Personalized PageRanking Vectors.

3.3.1 Seed Generation

We generate seeds by means of two different automatic procedures.

1. **AG**: We start at the *quality synset* of WordNet and iterate over WordNet relations following the original Q-WordNet method described in Agerri and García-Serrano (2010).

2. **TL**: We take a short manually created list of 14 positive and negative words (Turney and Littman, 2003) and iterate over WordNet using five relations: *antonymy, similarity, derived-from, pertains-to and also-see.*

The **AG** method starts the propagation from the attributes of the *quality synset* in WordNet. There are five noun quality senses in WordNet, two of which contain

	Synset Level							Word level						
		Positives			Negatives				Positives			Negatives		
Lexicon	**size**	**P**	**R**	**F**	**P**	**R**	**F**	**size**	**P**	**R**	**F**	**P**	**R**	**F**
Automatically created														
MSOL(ASL-GI)*	32706	.65	.45	.53	.58	.76	.66	76400	.70	.49	.58	.61	.79	.69
QWN	15508	.69	.53	.60	.62	.76	.68	11693	.64	.53	.58	.60	.70	.65
SWN	27854	.73	.57	.64	.65	.79	.71	38346	.70	.55	.62	.63	.77	.69
QWN-PPV-AG (s03_G1/w01_G1)	2589	.77	.63	.69	.69	.81	.74	5119	.68	.77	**.72**	.73	.64	**.68**
QWN-PPV-TL (s04_G1/w01_G1)	5010	.76	.66	**.70**	.70	.79	**.74**	4644	.68	.71	.69	.70	.67	.68
(Semi-) Manually created														
GI*	2791	.74	.57	.64	.65	.80	.72	3376	.79	.64	.71	.70	.83	.76
OF*	4640	.77	.61	.68	.68	.81	.74	6860	.82	.71	.76	.74	.84	.79
Liu*	4127	.81	.63	**.71**	.70	.85	**.76**	6786	.85	.74	**.79**	.77	.87	**.82**
SO-CAL*	4212	.75	.57	.64	.65	.81	.72	6226	.82	.70	.76	.74	.85	.79

Table 3.1: Evaluation of lexicons at document level using Bespalov's Corpus.

attribute relations (to adjectives). From the $quality_n^1$ synset the attribute relation takes us to $positive_a^1$, $negative_a^1$, $good_a^1$ and bad_a^1; $quality_n^2$ leads to the attributes $superior_a^1$ and $inferior_a^2$. The following step is to iterate through every WordNet relation collecting (i.e., annotating) those synsets that are accessible from the seeds. Both *AG* and *TL* methods to generate seeds rely on a number of relations to obtain a more balanced POS distribution in the output synsets. The output of both methods is a list of (assumed to be) positive and negative synsets. Depending on the number of iterations performed a different number of seeds to feed UKB is obtained. Seed numbers vary from 100 hundred to 10K synsets. Both seed creation methods can be applied to any WordNet, not only Princeton WordNet, as we show in section 3.4.

3.3.2 PPV generation

The second and last step to generate *QWN-PPV*(s) consists of propagating over a WordNet graph to obtain a Personalized PageRanking Vector (PPV), one for each polarity. This step requires:

1. A LKB projected over a graph.
2. A Personalized PageRanking algorithm which is applied over the graph.
3. Seeds to create contexts to start the propagation, either word or synsets.

Several undirected graphs based on WordNet 3.0 as represented by the MCR 3.0 (Agirre et al., 2012) have been created for the experimentation, which correspond to 4 main sets: (G1) two graphs consisting of every synset linked by the *synonymy* and *antonymy* relations; (G2) a graph with the nodes linked by every relation, including glosses; (G3) a graph consisting of the synsets linked by every relation except those that are linked by *antonymy*; finally, (G4) a graph consisting of the nodes related by every relation except the *antonymy* and *gloss* relations.

Using the (G1) graphs, we propagate from the seeds over each type of graph (synonymy and antonymy) to obtain two rankings per polarity. The graphs created in (G2), (G3) and (G4) are used to obtain two ranks, one for each polarity by propagating from the seeds. In all four cases the different polarity rankings have to be combined in order to obtain a final polarity lexicon: the polarity score *pol(s)* of a given synset s is computed by adding its scores in the positive rankings and subtracting its scores in the negative rankings. If *pol(s)* > 0 then s is included in the final lexicon as positive. If *pol(s)* < 0 then s is included in the final lexicon as negative. We assume that synsets with *null* polarity scores have no polarity and consequently they are excluded from the final lexicon.

The Personalized PageRanking propagation is performed starting from both synsets and words and using both *AG* and *TL* styles of seed generation, as explained in section 3.3.1. Combining the various possibilities will produce at least 6 different lexicons for each iteration, depending on which decisions are taken about which graph, seeds and word/synset to create the *QWN-PPV*(s). In fact, the experiments produced hundreds of lexicons, according to the different iterations for seed generation[1], but we will only refer to those that obtain the best results in the extrinsic evaluations.

With respect to the algorithm to propagate over the WordNet graph from the automatically created seeds, we use a Personalized PageRank algorithm (Agirre et al., 2014; Agirre and Soroa, 2009). The famous PageRank (Brin and Page, 1998) algorithm is a method to produce a rank from the vertices in a graph according to their relative structural importance. PageRank has also been viewed as the result of a Random Walk process, where the final rank of a given node represents the probability of a random walk over the graph which ends on that same node. Thus, if we take the created WordNet graph G with N vertices $v_1, \ldots, v_n$ and d_i as being the outdegree of node i, plus a $N \times N$ transition probability matrix M where $M_{ji} = 1/d_i$ if a link

[1]The total time to generate the final 352 QWN-PPV propagations amounted to around two hours of processing time in a standard PC.

	Synset Level							Word level						
		Positives			Negatives				Positives			Negatives		
Lexicon	**size**	**P**	**R**	**F**	**P**	**R**	**F**	**size**	**P**	**R**	**F**	**P**	**R**	**F**
Automatically created														
MSOL(ASL-GI)*	32706	.56	.37	.44	.76	.87	.81	76400	.67	.5	.57	.80	.89	.85
QWN	15508	.63	.22	.33	.73	.94	.83	11693	.58	.22	.31	.73	.93	.82
SWN	27854	.57	.33	.42	.75	.89	.81	38346	.55	.55	.55	.80	.8	.80
QWN-PPV-AG (w10_G3/s09_G4)	117485	.60	.63	**.62**	.83	.82	**.83**	144883	.65	.50	.57	.80	.88	.84
QWN-PPV-TL (s05_G4)	114698	.61	.58	.59	.82	.83	.83	144883	.66	.53	**.59**	.81	.88	**.84**
(Semi-) Manually created														
GI*	2791	.70	.32	.44	.76	.94	.84	3376	.71	.56	.62	.82	.90	.86
OF*	4640	.67	.37	**.48**	.77	.92	**.84**	6860	.75	.68	**.71**	.87	.90	**.88**
Liu*	4127	.67	.33	.44	.76	.93	.83	6786	.78	.45	.57	.79	.94	.86
SO-CAL*	4212	.69	.3	.42	.75	.94	.84	6226	.73	.53	.61	.81	.91	.86

Table 3.2: Evaluation of lexicons using *averaged ratio* on the MPQA 1.2_{test} Corpus.

from i to j exists and 0 otherwise, then calculating the PageRank vector over a graph G amounts to solve the following equation (3.1):

$$\mathbf{Pr} = cM\mathbf{Pr} + (1 - c)\mathbf{v} \tag{3.1}$$

In the traditional PageRank, vector $\mathbf{v}$ is a uniform normalized vector whose elements values are all $1/N$, which means that all nodes in the graph are assigned the same probabilities in case of a random walk. Personalizing the PageRank algorithm in this case means that it is possible to make vector $\mathbf{v}$ non-uniform and assign stronger probabilities to certain nodes, which would make the algorithm to propagate the initial importance of those nodes to their vicinity. Following Agirre et al. (2014), in our approach this translates into initializing vector $\mathbf{v}$ with those senses obtained by the seed generation methods described above in section 3.3.1. Thus, the initialization of vector $\mathbf{v}$ using the seeds allows the Personalized propagation to assign greater importance to those synsets in the graph identified as being positive and negative, which resuls in a PPV with the weigths skewed towards those nodes initialized/personalized as positive and negative.

3.4 Evaluation

Previous approaches have provided intrinsic evaluation (Mohammad et al., 2009; Rao and Ravichandran, 2009; Baccianella et al., 2010) using manually annotated

resources such as the General Inquirer (Stone et al., 1966) as gold standard. To facilitate comparison, we also provide such evaluation in section 3.4.3. Nevertheless, and as demonstrated by the results of the extrinsic evaluations, we believe that polarity lexicons should in general be evaluated *extrinsically*. After all, any polarity lexicon is as good as the results obtained by using it for a particular Sentiment Analysis task.

Our goal is to evaluate the polarity lexicons simplifying the evaluation parameters to avoid as many external influences as possible on the results. We compare our work with most of the lexicons reviewed in section 3.2, both at synset and word level, both manually and automatically generated: General Inquirer (**GI**), Opinion Finder (**OF**), **Liu**, Taboada *et al.'s* (**SO-CAL**), Agerri and García-Serrano's (**QWN**), Mohammad *et al's* (**MSOL(ASL-GI)**) and SentiWordNet 3.0 (**SWN**). The results presented in section 3.4.2 show that *extrinsic* evaluation is more meaningful to determine the adequacy of a polarity lexicon for a specific Sentiment Analysis task.

3.4.1 Datasets and Evaluation System

Three different corpora were used: Bespalov *et al.'s* (2011) and MPQA (Riloff and Wiebe, 2003a) for English, and HOpinion[2] in Spanish. In addition, we divided the corpus into two subsets (75% development and 25% test) for applying our ratio system for the phrase polarity task too. Note that the development set is only used to set up the polarity classification task, and that the generation of *QWN-PPV* lexicons is unsupervised.

For Spanish we tried to reproduce the English settings with Bespalov's corpus. Thus, both development and test sets were created from the HOpinion corpus. As it contains a much higher proportion of positive reviews, we created also subsets which contain a balanced number of positive and negative reviews to allow for a more meaningful comparison than that of table 3.6. Table 3.3 shows the number of documents per polarity for Bespalov's, MPQA 1.2 and HOpinion.

We report results of 4 extrinsic evaluations or tasks, three of them based on a simple *ratio average system*, inspired by Turney (2002), and another one based on (Mohammad et al., 2009). We first implemented a simple *average ratio classifier* which computes the average ratio of the polarity words found in document d:

$$polarity(d) = \frac{\sum_{w \in d} pol(w)}{|d|} \tag{3.2}$$

[2] `http://clic.ub.edu/corpus/hopinion`

Corpus	**POS docs**	**NEG docs**	**Total**
Bespalov$_{dev}$	23,112	23,112	46,227
Bespalov$_{test}$	10,557	10,557	21,115
MPQA 1.2$_{dev}$	2,315	5,260	7,575
MPQA 1.2$_{test}$	771	1,753	2,524
MPQA 1.2$_{total}$	3,086	7,013	10,099
HOpinion_Balanced$_{dev}$	1,582	1,582	3,164
HOpinion_Balanced$_{test}$	528	528	1,056
HOpinion$_{dev}$	9,236	1,582	10,818
HOpinion$_{test}$	3,120	528	3,648

Table 3.3: Number of positive and negative documents in train and test sets.

	Synset Level							**Word level**						
		Positives			**Negatives**				**Positives**			**Negatives**		
Lexicon	**size**	**P**	**R**	**F**	**P**	**R**	**F**	**size**	**P**	**R**	**F**	**P**	**R**	**F**
Automatically created														
MSOL(ASL-GI)*	32706	.52	.48	.50	.85	.62	.71	76400	.68	.56	.62	.82	.86	.84
QWN	15508	.50	.36	.42	.84	.32	.46	11693	.45	.49	.47	.78	.51	.61
SWN	27854	.50	.45	.47	.85	.48	.61	38346	.49	.52	.50	.78	.68	.73
QWN-PPV-AG (s09_G3/w02_G3)	117485	.59	.67	**.63**	.85	.78	**.82**	147194	.64	.64	.64	.84	.83	.83
QWN-PPV-TL (w02_G3/s06_G3)	117485	.59	.57	.58	.82	.81	.81	147194	.63	.67	**.65**	.85	.81	**.83**
(Semi-) Manually created														
GI*	2791	.60	.40	.47	.91	.38	.54	3376	.70	.60	.65	.93	.52	.67
OF*	4640	.63	.42	**.50**	.93	.46	**.62**	6860	.75	.71	**.73**	.95	.66	**.78**
Liu*	4127	.65	.36	.47	.94	.45	.60	6786	.78	.49	.60	.97	.61	.75
SO-CAL*	4212	.65	.37	.47	.92	.45	.60	6226	.73	.57	.64	.96	.59	.73

Table 3.4: Evaluation of lexicons at phrase level using Mohammad *et al.'s* (2009) method on MPQA 1.2$_{total}$ Corpus.

where, for each polarity, *pol(w)* is 1 if w is included in its polarity lexicon and 0 otherwise. Documents that reach a certain threshold are classified as positive, and otherwise as negative. To setup an evaluation enviroment as fair as possible for every lexicon, the threshold is optimised by maximising accuracy over the development data.

Second, we implemented a phrase polarity task identification as described by Mohammad et al. (2009). Their method consists of: (i) if any of the words in the

target phrase is contained in the negative lexicon, then the polarity is negative; (ii) if none of the words are negative, and at least one word is in the positive lexicon, then is positive; (iii) the rest are not tagged.

We chose this very simple polarity estimators because our aim was to minimize the role other aspects play in the evaluation and focus on how, other things being equal, polarity lexicons perform in a Sentiment Analysis task. The *average ratio* is used to present results of tables 3.1 (with Bespalov corpus) 3.5 and 3.6 (both with HOpinion), where as Mohammad *et al.'s* is used to report results in tables 3.2 (with MPQA 1.2_{test}) and 3.4 (with MPQA 1.2_{total}). Mohammad *et al.'s* (2009) testset based on MPQA 1.1 is smaller, but both MPQA 1.1 and 1.2 are hugely skewed towards negative polarity (30% positive vs. 70% negative).

All datasets were POS tagged and Word Sense Disambiguated using FreeLing (Padró and Stanilovsky, 2012; Agirre and Soroa, 2009). Having word sense annotated datasets gives us the opportunity to evaluate the lexicons both at word and sense levels. For the evaluation of those lexicons that are synset-based, such as *QWN-PPV* and SentiWordNet 3.0, we convert them from senses to words by taking every word or variant contained in each of their senses. Moreover, if a lemma appears as a variant in several synsets the most frequent polarity is assigned to that lemma.

With respect to lexicons at word level, we take the most frequent sense according to WordNet 3.0 for each of their positive and negative words. Note that the latter conversion, for synset based evaluation, is mostly done to show that the evaluation at synset level is harder independently of the quality of the lexicon evaluated.

3.4.2 Results

Although tables 3.1, 3.2 and 3.4 also present results at synset level, it should be noted that the only polarity lexicons available to us for comparison at synset level were Q-WordNet (Agerri and García-Serrano, 2010) and SentiWordNet 3.0 (Baccianella et al., 2010). *QWN-PPV-AG* refers to the lexicon generated starting from **AG**'s seeds, and *QWN-PPV-TL* using **TL**'s seeds as described in section 3.3.1. Henceforth, we will use *QWN-PPV* to refer to the overall method presented in this paper, regardless of the seeds used.

For every *QWN-PPV* result reported in this section, we have used every graph described in section 3.3.2. The configuration of each *QWN-PPV* in the results specifies which seed iteration is used as the initialization of the Personalized PageRank algorithm, and on which graph. Thus, QWN-PPV-TL (s05_G4) in table

3.2 means that the 5th iteration of synset seeds was used to propagate over graph G4. If the configuration were (w05_G4) it would have meant 'the 5th iteration of word seeds were used to propagate over graph G4'. The simplicity of our approach allows us to generate many lexicons simply by projecting a LKB over different graphs.

The lexicons marked with an asterisk denote those that have been converted from word to senses using the most frequent sense of WordNet 3.0. We would like to stress again that the purpose of such word to synset conversion is to show that SA tasks at synset level are harder than at word level. In addition, it should also be noted that in the case of SO-CAL (Taboada et al., 2011b), we have reduced what is a graded lexicon with scores ranging from 5 to -5 into a binary one.

Table 3.1 shows that (at least partially) manually built lexicons obtain the best results on this evaluation. It also shows that *QWN-PPV* clearly outperforms any other automatically built lexicons. Moreover, manually built lexicons suffer from the evaluation at synset level, obtaining most of them lower scores than *QWN-PPV*, although Liu's (Hu and Liu, 2004b) still obtains the best results. In any case, for an unsupervised procedure, *QWN-PPV* lexicons obtain very competitive results with respect to manually created lexicons and is the best among the automatic methods. It should also be noted that the best results of *QWN-PPV* are obtained with graph G1 and with very few seed iterations.

Table 3.2 again sees the manually built lexicons performing better although overall the differences are lower with respect to automatically built lexicons. Among these, *QWN-PPV* again obtains the best results, both at synset and word level, although in the latter the differences with MSOL(ASL-GI) are not large. Finally, table 3.4 shows that *QWN-PPV* again outperforms other automatic approaches and is closer to those have been (partially at least) manually built. In both MPQA evaluations the best graph overall to propagate the seeds is G3 because this type of task favours high recall.

We report results on the Spanish HOpinion corpus in tables 3.5 and 3.6. Mihalcea(f) is a manually revised lexicon based on the automatically built Mihalcea(m) (Pérez-Rosas et al., 2012). ElhPolar (Saralegi and San Vicente, 2013) is semi-automatically built and manually corrected. SO-CAL is built manually. SWN and QWN-PPV have been built via the MCR 3.0's ILI by applying the synset to word conversion previously described on the Spanish dictionary of the MCR. The results for Spanish at word level in table 3.6 show the same trend as for English: *QWN-PPV* is the best of the automatic approaches and it obtains competitive although not as good as the best of the manually created lexicons (ElhPolar). Due

		Positives			Negatives		
Lexicon	**size**	**P**	**R**	**F**	**P**	**R**	**F**
Automatically created							
SWN	27854	.87	.99	.93	.70	.16	.27
QWN-PPV-AG (wrd01_G1)	3306	.86	.00	.92	.67	.01	.02
QWN-PPV-TL (s04_G1)	5010	.89	.96	**.93**	.58	.30	**.39**

Table 3.5: Evaluation of Spanish lexicons using the full HOpinion corpus at synset level.

to the disproportionate number of positive reviews, the results for the negative polarity are not useful to draw any meaningful conclusions. Thus, we also performed an evaluation with HOpinion Balanced set as listed in table 3.3.

		Positives			Negatives		
Lexicon	**size**	**P**	**R**	**F**	**P**	**R**	**F**
Automatically created							
Mihalcea(m)	2496	.86	.00	.92	.00	.00	.00
SWN	9712	.88	.97	.92	.55	.19	.28
QWN-PPV-AG (s11_G1)	1926	.89	.97	.93	.59	.26	.36
QWN-PPV-TL (s03_G1)	939	.89	.98	**.93**	.71	.26	**.38**
(Semi-) Manually created							
ElhPolar	4673	.94	.94	**.94**	.64	.64	**.64**
Mihalcea(f)	1347	.91	.96	.93	.61	.41	.49
SO-CAL	4664	.92	.96	.94	.70	.51	.59

Table 3.6: Evaluation of Spanish lexicons using the full HOpinion corpus at word level.

The results with a balanced HOpinion, not shown due to lack of space, also confirm the previous trend: *QWN-PPV* outperforms other automatic approaches but is still worse than the best of the manually created ones (ElhPolar).

3.4.3 Intrinsic evaluation

To facilitate intrinsic comparison with previous approaches, we evaluate our automatically generated lexicons against GI following Mohammad et al.'s (2009) method. For each *QWN-PPV* lexicon shown in previous extrinsic evaluations,

we compute the intersection between the lexicon and GI, and evaluate the words in that intersection. Table 3.7 shows results for the best-performing QWN-PPV lexicons (both using AG and TL seeds) in the extrinsic evaluations at word level of tables 3.1 (first two rows), 3.2 (rows 3 and 4) and 3.4 (rows 5 and 6). We can see that QWN-PPV lexicons systematically outperform SWN in number of correct entries. *QWN-PPV-TL* lexicons obtain 75.04% of correctness on average. The best performing lexicon contains up to 81.07% of correct entries. Note that we did not compare the results with MSOL(ASL-GI) because it uses the GI as seeds.

Lexicon	**∩ wrt. GI**	**Acc.**	**Pos**	**Neg**
SWN	2,755	.74	.76	.73
QWN-PPV-AG (w01_G1)	849	.71	.68	.75
QWN-PPV-TL (w01_G1)	713	.78	.80	.76
QWN-PPV-AG (s09_G4)	3,328	.75	.75	.77
QWN-PPV-TL (s05_G4)	3,333	**.80**	**.84**	**.77**
QWN-PPV-AG (w02_G3)	3,340	.74	.71	.77
QWN-PPV-TL (s06_G3)	3,340	.77	.79	.77

Table 3.7: Accuracy QWN-PPV lexicons and SWN with respect to the GI lexicon.

3.4.4 Discussion

QWN-PPV lexicons obtain the best results among the evaluations for English and Spanish. Furthermore, across tasks and datasets *QWN-PPV* provides a more consistent and robust behaviour than most of the manually-built lexicons apart from OF. The results also show that for a task requiring high recall the larger graphs, e.g. G3, are preferable, whereas for a more balanced dataset and document level task smaller G1 graphs perform better.

These are good results considering that our method to generate *QWN-PPV* is simpler, more robust and adaptable than previous automatic approaches. Furthermore, although also based on a Personalized PageRank application, it is much simpler than SentiWordNet 3.0, consistently outperformed by *QWN-PPV* on every evaluation and dataset. The main differences with respect to SentiWordNet's approach are the following: (i) the seed generation and training of 7 supervised classifiers corresponds in *QWN-PPV* to only one simple step, namely, the automatic generation of seeds as explained in section 3.3.1; (ii) the generation of *QWN-PPV* only requires a LKB's graph for the Personalized PageRank propagation, no

disambiguated glosses; (iii) the graph they use to do the propagation also depends on disambiguated glosses, not readily available for any language.

The fact that *QWN-PPV* is based on already available WordNets projected onto simple graphs is crucial for the robustness and adaptability of the *QWN-PPV* method across evaluation tasks and datasets: Our method can quickly create, over different graphs, many lexicons of diffent sizes which can then be evaluated on a particular polarity classification task and dataset. Hence the different configurations of the *QWN-PPV* lexicons, because for some tasks a G3 graph with more AG/TL seed iterations will obtain better recall and viceversa. This is confirmed by the results: the tasks using MPQA seem to clearly benefit from high recall whereas the Bespalov's corpus has overall, more balanced scores. This could also be due to the size of Bespalov's corpus, almost 10 times larger than MPQA 1.2.

The experiments to generate Spanish lexicons confirm the trend showed by the English evaluations: Lexicons generated by *QWN-PPV* consistenly outperform other automatic approaches, although some manual lexicon is better on a given task and dataset (usually a different one). Nonetheless the Spanish evaluation shows that our method is also robust across languages as it gets quite close to the manually corrected lexicon of Mihalcea(full) (Pérez-Rosas et al., 2012).

The results also confirm that no single lexicon is the most appropriate for any SA task or dataset and domain. In this sense, the adaptability of *QWN-PPV* is a desirable feature for lexicons to be employed in SA tasks: the unsupervised *QWN-PPV* method only relies on the availability of a LKB to build hundreds of polarity lexicons which can then be evaluated on a given task and dataset to choose the best fit. If not annotated evaluation set is available, G3-based propagations provide the best recall whereas the G1-based lexicons generate are less noisy. Finally, we believe that the results reported here point out to the fact that intrinsic evaluations are not meaningful to judge the adequacy a polarity lexicon for a specific SA task.

3.5 Concluding Remarks

This paper presents an unsupervised dictionary-based method *QWN-PPV* to automatically generate polarity lexicons. Although simpler than similar automatic approaches, it still obtains better results on the four extrinsic evaluations presented. Because it only depends on the availability of a LKB, we believe that this method can be valuable to generate on-demand polarity lexicons for a given language when

not sufficient annotated data is available. We demonstrate the adaptability of our approach by producing good performance polarity lexicons for different evaluation scenarios and for more than one language.

Further work includes investigating different graph projections of WordNet relations to do the propagation as well as exploiting synset weights. We also plan to investigate the use of annotated corpora to generate lexicons at word level to try and close the gap with those that have been (at least partially) manually annotated.

The *QWN-PPV* lexicons and graphs used in this paper are publicly available (under CC-BY license): *http://adimen.si.ehu.es/web/qwn-ppv*. The *QWN-PPV* tool to automatically generate polarity lexicons given a WordNet in any language will soon be available in the aforementioned URL.

Acknowledgements

This work has been supported by the OpeNER FP7 project under Grant No. 296451, the FP7 NewsReader project, Grant No. 316404 and by the Spanish MICINN project SKATER under Grant No. TIN2012-38584-C06-01.

CHAPTER 4

Method Comparison

Polarity Lexicon Building: to what Extent Is the Manual Effort Worth?

Iñaki San Vicente, Xabier Saralegi

Elhuyar Fundazioa

Polarity lexicons are a basic resource for analyzing the sentiments and opinions expressed in texts in an automated way. This paper explores three methods to construct polarity lexicons: translating existing lexicons from other languages, extracting polarity lexicons from corpora, and annotating sentiments Lexical Knowledge Bases. Each of these methods require a different degree of human effort. We evaluate how much manual effort is needed and to what extent that effort pays in terms of performance improvement. Experiment setup includes generating lexicons for Basque, and evaluating them against gold standard datasets in different domains. Results show that extracting polarity lexicons from corpora is the best solution for achieving a good performance with reasonable human effort.

Published in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, may 2016. ISBN 978-2-9517408-9-1.

4.1 Introduction

Research effort on the sentiment analysis field has seen exponentially increased in the last years, due to its applicability in areas such as VTIC (Technological Surveillance / Competitive Intelligence), marketing or reputation management. One of the main resources of sentiment analysis systems are the polarity lexicons, list of words with prior polarities. Much research has been done on methods for building such methods automatically due to the high cost of manually created lexicons. Then again, automatic methods often produce noisy resources.

Very little work has been done on polarity lexicons for Basque, as is the case for other less resourced languages. Thus, when facing the task of creating such a resource, the doubt arised. Is it worth to make a great manual annotation effort? How much is the gain we obtain by manually annotating polarity words over automatically built polarity lexicons?

This paper compares three strategies for building a polarity lexicon for a less resourced language. We assumed that for languages of this type the availability of parallel corpora, MT systems and polarity-annotated data is very limited, and we avoided using such resources. We measured the time cost of the manual effort and the gain it brings in terms of accuracy in an extrinsic evaluation. This experiment was carried out for Basque.

4.2 State of the Art

Polarity lexicons are key resource on sentiment analysis systems. We can group the methods for polarity lexicon building proposed in the literature into three main approaches: manually constructed lexicons (Stone et al., 1966), corpus-based methods (Hatzivassiloglou and McKeown, 1997; Mihalcea et al., 2007) and methods that rely on Lexical Knowledge Bases (LKB) (Kamps et al., 2004; Liu and Singh, 2004; Kim and Hovy, 2004b) For major languages there are well known manually constructed lexicons, such as, General Inquirer(Stone et al., 1966), OpinionFinder(Wilson et al., 2005), or SO-CAL (Taboada et al., 2011a). Due to the fact that a great human effort is needed to build such resources, some of them are semi-automatically constructed, and manually corrected afterwards. In this line of work, some researchers explore the possibility of using resources already existing in another language (e.g., lexicons, and/or annotated corpora corpora). (Mihalcea et al., 2007) and (Perez-Rosas et al., 2012) analyze the approach of translating English resources into Romanian and

Spanish, respectively. However, only a small portion of the translated lexicon entries maintain the correct polarity. The need to treat ambiguous translations becomes clear.

Corpus-based methods require some sort of polarity annotation to construct the lexicons. We can find two main approaches in this group: i) starting from a small list of words with known polarity, find words in a corpus that are semantically close by means of distributional methods (Turney and Littman, 2003), and ii) Based on a corpus that has polarity annotations at document or sentence level, create list of words most related to either positive or negative annotations (Saralegi and San Vicente, 2012).

Finally, the main idea behind LKB-based methods is to propagate to new words the polarity of a small list of seed words with known polarities, by making use of relations between concepts the LKB offers. Propagating polarity through graphs representing the semantic relations existing in WordNet (WN) (Fellbaum, 1998) is a well known strategy (Esuli and Sebastiani, 2006; San Vicente et al., 2014).

With respect to the specific case of Basque, we have found two polarity lexicons in the literature. The NRC Word-Emotion association lexicon, constructed in a crowdsourcing annotation effort, was translated using Google Translate to Basque (NRC_{eu}) (Mohammad and Turney, 2013). The second lexicon is MLSenticon (Cruz et al., 2014), which is an LKB-based lexicon generated in a similar way to SentiWordNet (Esuli and Sebastiani, 2006).

4.3 Lexicon Building methods.

Our aim is to compare three methods for polarity lexicon building which require a different degree of human edition: (i) Translating lexicons in other language into our language; (ii) extracting automatically polarity words from corpora; eta (iii) annotating the polarity of the words in an LKB.

4.3.1 Projection

Projecting polarity lexicons from other languages by means of bilingual dictionaries seems like a direct way to create a lexicon in our language. However, this approach has to deal with the problems derived from the translation process: ambiguous translations and changes in the polarity of the target words.

Spanish lexicon $ElhPolar_{es}$ (Saralegi and San Vicente, 2013) has been translated by means of the Elhuyar Spanish-Basque dictionary[1] (173,931 translation pairs). For each Spanish entry in the lexicon, the first 5 translations are included in the translated lexicon Lex_{pr}.

Lex_{pr} has been initially reviewed by a native speaker correcting the polarity of each word. 4.3.4 offers details on the cost of this correction effort. Furthermore, a second reference by another annotator was later carried out on part of Lex_{pr}. Details about this second annotation effort are given in section 4.4.

The corrected lexicon contains 5,335 entries, 1,938 positive and 3,397 negative, very similar numbers to its original Spanish version.

	#entry	**#positive**	**#negative**
$ElhPolar_{es}$	5.195	1.892	3.303
Lex_{pr}	11.413	4.934	6.479

Table 4.1: ElhPolar source and translated lexicons' statistics.

4.3.2 Corpus-based lexicons

The second approach is based on the idea that words that tend to appear in texts with a certain polarity (positive or negative) are good representatives of that polarity. Usually association measures (AM) are used to find salient words in corpora (Kilgarriff, 2001).

Ideally, we would use a corpus with polarity annotations, which we could divide into positive and negative subparts. Unfortunately, no such resource exists for Basque and many other less resourced languages. As a solution, we adopted a semi-automatic approach relying on a corpus including subjective and objective documents (Saralegi et al., 2013). Such a corpus can be built in an easy way from a newspaper corpus taking as subjective documents opinion articles and as objective event news.

Using the Loglikelihood ratio (LLR) (Dunning, 1993) we obtained the ranking of the most salient words in the subjective part with respect to the rest of the corpus. The top 5,000 subjective words were manually checked by a single annotator. The corrected lexicon (Lex_c) contains 1.659 entries (959 negative and 691 positive).

[1] http://hiztegiak.elhuyar.eus

This method ranks a lot polar word candidates among the first positions because subjectivity highly correlates to polar words.

4.3.3 LKB-based lexicons

Using the semantic relations represented in LKBs in order to construct polarity lexicons is a widespread strategy in the literature. In our case, we apply the method presented in (San Vicente et al., 2014) for generating basque polarity lexicons. *Q-WordNet as Personalized PageRanking Vector* (QWN-PPV) represents the concepts and the semantic relations between them stored in a WN like LKB over a graph. The method propagates the polarity of an initial set of words by applying the so-called Personalized PageRank algorithm on a LKB. We use the UKB (Agirre and Soroa, 2009) implementation of the algorithm.

The Basque WN (Pociello et al., 2011) is small compared to others. Thus, we chose to use MCR (Agirre et al., 2012) as LKB. Because it connects WNs for several languages including Basque, we can take advantage of a number of semantic relations existing in larger WNs which offer a bigger chance to propagate polarity information. Two graph representations are used, one including synonymy relations and another antonymy relations. We chose this graph representation because it creates higher quality propagations, although the limited number of relations results on smaller lexicons.

The lexicon produced with this approach ($Lex_{qwn-ppv}$) contains 1.132 entries, 565 positive and 567 negative.

The settings used in this work for QWN-PPV are derived from the experiments carried out in (San Vicente et al., 2014).

4.3.4 Correction effort

Usually, the main problem of the manual effort is its high cost. In this work we have measured the annotation effort required to correct the lexicons. As an indicator of that effort we have used what we call production rate. We understand production rate as the number of words added to our lexicon per minute.

Projection

Altogether, a single annotator needed 36 hours to correct the Basque projected lexicon Lex_{pr}. That means that the correction rate was 5,3 word/minute. As general

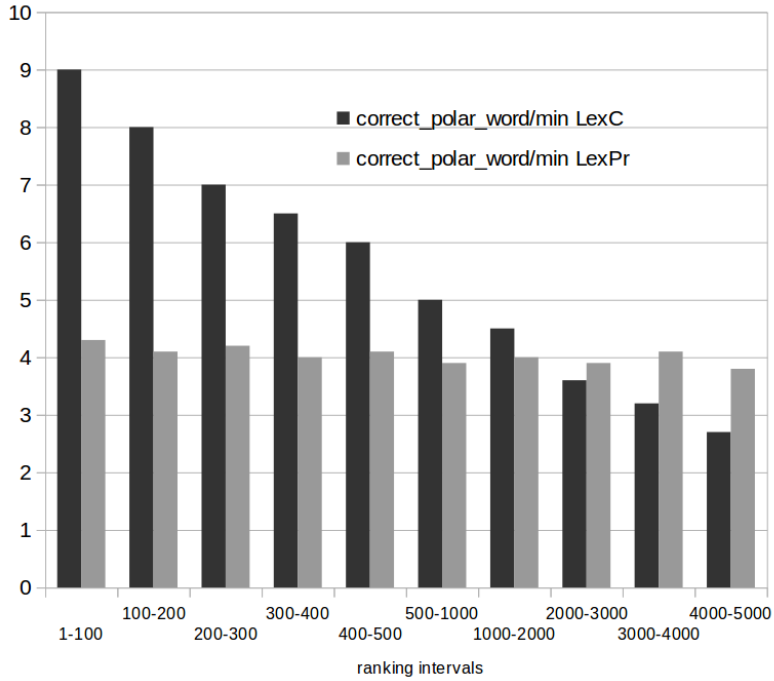


Figure 4.1: Correction speed and productivity data for Lex_{pr} and Lex_c.

remark, we can say that the correction requires a great manual effort, because the dictionary-based translation selects many unusual translations (rarely used words), which leads annotators to consult frequently dictionaries and corpora.

Corpus-based lexicon

In contrast to the translation approach, the annotator must decide the polarity of a word without any prior information on this regard, but, on the other hand, since the words are extracted from a corpus by means of LLR, the list contains more frequently used words. Hence, it is easier to annotate the polarity of common words as dictionaries and corpora are not so frequently needed. Overall, 10 hours were needed to annotate the polarity of the 5,000 candidate list. This means a correction rate of 8,3 word/minute.

Figure 4.1 shows the average production rates of the annotation process, for the various candidate ranking intervals. The higher production rates achieved for the first ranked candidates in the corpus-based method (correct_polar_word/min Lex_C),

is due to those candidates being most frequent words. The deeper we go into the ranking, the more unusual words appear, and hence the correction speed is reduced. Also, the higher production rate observed for the corpus-based method indicates that indeed LLR surfaces polar words, having a higher density in the first positions of the ranking. The top-ranking words are those which have most association degree with subjective corpus. For comparison between projection and corpus-based methods, only data for the first 5,000 candidates is shown.

4.3.5 Second reference

A single reference may not be fully trustworthy, and so we introduced a second reference for both projected and corpus-based lexicons. Due to the time constraints, we asked the second annotator only to review those words in the intersections between the lexicons and the datasets evaluated. Disagreements were resolved by discussion. This second annotation allows us to measure the improvement we can gain with that extra effort, as will be explained in section 4.4. Table 4.2 shows the number of lemmas annotated on this second annotation, inter-annotator agreement (Cohen's Kappa κ value) data for positive (+) and negative (-) words and the time spent on discussing the disagreement cases.

Lexicon	**#Lemmas annotated**	κ +	κ -	**Disagree-ments**	**Discussion time (min)**
Lex_{pr}	599	0.624	0.765	80	65
Lex_C	542	0.747	0.835	56	40

Table 4.2: Statistics for the second annotation effort.

4.4 Evaluation

In order to evaluate the adequacy of the generated lexicons we set up a binary polarity classification task (positive vs. negative). As there is no corpus with gold annotations, we have generated two small datasets manually annotated at sentence level. Section 4.4 give details about those datasets.

Classifier

We implement a simple average polarity ratio classifier. There are two reasons to choose such classifier: on the one hand, the lack of and annotated corpus prevents us from using supervised classifiers, and on the other, our aim is to minimize the role other aspects play in the evaluation and focus on how, other things being equal, polarity lexicons perform in a Sentiment Analysis task. The *average ratio classifier* computes the average ratio of the polarity words found in document d:

$$Pol(d) = \frac{\sum_{w \in d} pol(w)}{\#w} \tag{4.1}$$

where, for each word w, *pol(w)* is the polarity of the word in the lexicon ($1 = positive, -1 = negative$) or 0 if the word is missing. If $Pol(d) > 0$ d is classified as positive, and otherwise as negative.

Evaluated lexicons

Our aim is to evaluate to what extent manual effort brings improvement. Overall we include 11 lexicons in the evaluation. For both Projection and Corpus-based lexicons 3 lexicons are evaluated, one for each of the annotators (Rows starting "AnnotX" in table 4.4) and a third generated from the consensus of those annotations (Rows starting "Consens" in table 4.4). In addition, the projected lexicon before manual annotation is included as a baseline. The LKB based lexicon provides comparison with a fully automatic method. The combination of both corpus-based and projected lexicons annotated represents what the greatest manual effort can achieve. Lastly, for the sake of comparison, although we didn't build them, we include the two publicly available polarity lexicons for Basque found in the literature: NRC_{eu} and MLSenticon.

Test datasets

Two test-sets were compiled from different sources: One from the news domain, composed of newspaper articles, and another one from music and film reviews. Overall 224 sentences were gathered and manually annotated as positive and negative (see table 4.3). Neutral polarity sentences were discarded.

Domain	Positive	Negative	Overall
Music&Film reviews	%75.58	%24.42	86
News	%25.36	%74.64	138
Overall	%44.64	%55.36	224

Table 4.3: Test datasets estatistics.

4.4.1 Results

Lexicon	News			Music&Films			Overall		
	Acc.	**Fpos**	**Fneg**	**Acc.**	**Fpos**	**Fneg**	**Acc.**	**Fpos**	**Fneg**
Projection									
Lex_{pr}	0.63	0.41	0.73	0.63	0.69	0.53	0.63	0.57	0.68
Annot1-Lex_{pr}	0.80	0.61	0.87	0.67	0.75	0.55	0.75	0.69	0.80
Annot2-Lex_{pr}	0.78	0.61	0.85	**0.76**	0.81	0.67	0.77	0.72	0.81
ConsensLex_{pr}	**0.86**	0.68	0.91	0.70	0.75	0.62	**0.79**	0.72	0.84
Corpus-based									
Annot1-Lex_c	0.77	0.56	0.84	0.79	0.85	0.64	**0.78**	0.74	0.80
Annot2-Lex_c	0.75	0.48	0.84	0.74	0.81	0.61	0.75	0.69	0.79
Consens-Lex_c	**0.78**	0.56	0.86	**0.80**	0.86	0.67	**0.79**	0.75	0.82
Automatic									
$Lex_{qwn-ppv}$	0.67	0.21	0.79	0.55	0.68	0.20	0.63	0.53	0.69
Combination									
ConsensLex_{c+pr}	**0.88**	0.74	0.92	**0.83**	0.87	0.73	**0.86**	0.82	0.88
External									
NRC_{eu}	0.62	0.29	0.74	0.47	0.51	0.41	0.56	0.41	0.65
$MLSenticon$	0.65	0.37	0.76	0.55	0.60	0.48	0.61	0.50	0.68

Table 4.4: Evaluation results for the various lexicons on the test datasets.

Table 4.4 presents the results obtained by the various lexicons over the test datasets. Accuracy (Acc.) and F-score values per category (Fpos/Fneg) are reported. Corpus-based lexicon achieves the best results across all datasets. As expected manually corrected lexicons perform better that the automatically generated lexicon.

Overall, results show corpus-based lexicons obtain very similar results to those of the translated lexicons, with much less human effort. Furthermore corpus-based lexicons' performance is far better in the Music&Film review domain.

Also, results show that a second annotation and the following discussion does indeed improve the quality of the lexicon in terms of accuracy. This of course means

a greater annotation effort.

As an upper bound, the combination of the translated and corpus-based lexicons obtains the best results overall, although it also means the greatest annotation effort.

The performance of the automatically built LKB-based lexicon is far from the manually corrected lexicons, although its performance is similar to that of Lex_{pr}, the other completely automatic lexicon in the evaluation. Moreover, Basque WN suffers from a severe lack of information on adjectives. As adjectives are important for polarity detection, a better coverage would improve the lexicons generated with this strategy.

With respect to external lexicons, NRC_{eu} obtains modest results. There are to main reasons that led to its poor performance. The lexicon contains some incorrect entries and many of the entries correspond to word forms instead of lemmas. This is probably a side effect of the automatic translation. MLSenticon's results are very close to our own automatic Lexicon $Lex_{qwn-ppv}$. This is not surprising, since they both rely in a similar method and use MCR to obtain Basque lemmas.

4.5 Discussion and Conclusions

This paper explores three methods to build polarity lexicons from scratch. The adequacy of those methods has been evaluated on a polarity classification task over data from two different domains.

Semi-automatic corpus-based generation of polarity lexicons would be an adequate approach for scenarios where time for manual effort is limited. The manual effort required in this strategy is not very costly (10 hours). Even if the lexicon is not very large, the fact that it is corpus-based guaranties that most used polar words will be present.

For the scenarios where the accuracy is critical the combination of both projection and corpus-based strategies with at least two annotators would be desirable for building the polarity lexicon.

We plan to extend this research by constructing new polarity annotated datasets. This will allow us, on the one hand, to evaluate our resources using a machine-learning approach, which would be the first ML sentiment analysis system for Basque; and, on the other, new datasets would provide resources to generate new lexicons. Finally, repeating the experiments with other languages would add robustness to the contribution of this paper.

Acknowledgments

This work has been supported by Basque Government Elkartek program, in the framework of the Elkarola project (grants no. KK-2015/00098, KK-2016/00087).

PART II

ANALYSIS OF SOCIAL MEDIA

II Analysis of Social Media

This part covers the work done in order to pre-process data coming from social media in order to apply polarity classification to that data. Although we have specifically worked with Twitter, the findings could be to some extent be valid for other social media such as Facebook or Instagram. A Study of User behaviour across social media by (Lim et al., 2015) reports that more than 95% of users from other social media (Facebook, Google+, Instagram, Tumblr, Flickr and Youtube) are also active in Twitter, and define twitter as a sink destination, revealing that 54% of the messages posted in other social networks are also shared in Twitter.

Two main challenges have been addressed in this part: a) language identification of less resourced language messages in Twitter; and b) Tweet normalization. There are some other preprocesses that may be applied to Twitter messages related to Sentiment Analysis, such as entity standardization (usernames, #hashtags, urls) and emoji mapping. Those processes are covered in Part III.

The TweetLID shared task (Zubiaga et al., 2016) (chapter 5) provided a benchmark for language identification in Twitter, covering the official languages of the Iberian peninsula. This covered the problematic of similar languages (Galician-Portuguese, Catalan-Spanish) and that of less resourced languages (Basque, Galician, Catalan). The author of this thesis was part of the organizing committee, and took active part on the design of the task, annotation of the data, coordination of the task and also in the evaluation of the submitted systems.

TweetNorm (Alegria et al., 2015) (chapter 6) established the first benchmark for Spanish microtext normalization. The paper presents the dataset created, the evaluation campaign organized and the analysis of the systems submitted.

Chapter 7 gives the details of the submitted system (Saralegi and San Vicente, 2013b). The system developed relies on a series of heuristics and normalization

resources for generating standard form candidates for OOV words, and then makes use of a language model in order to find the best candidates.

The findings in (Saralegi and San Vicente, 2013b) and (Alegria et al., 2015) served as the basis to implement the microtext normalization module that is integrated in EliXa, as we will see in chapter 10.

Finally, we list the publications included in this part, describing also the contribution of the author of this thesis in each of the papers.

- Arkaitz Zubiaga, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. Tweetlid: a benchmark for tweet language identification. *Language Resources and Evaluation*, 50(4):729–766, Dec 2016. ISSN 1574-0218

  Contribution to the paper: Iñaki San Vicente was part of the organizing committee. Took part on the annotation of the datasets, and created the evaluation scripts for the task. Arkaitz Zubiaga was the main coordinator of the task.

- Iñaki Alegria, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. Tweetnorm: a benchmark for lexical normalization of spanish tweets. *Language Resources and Evaluation*, 49(4):883–905, Dec 2015. ISSN 1574-0218

  Contribution to the paper: Iñaki San Vicente was part of the organizing committee. He took part in the coordination of the shared task, as well as in the evaluation of the submissions. Although he took part to a certain extent in the creation of the guidelines, he did not participate in the annotation. The reason was that he also developed a system that participated in the shared task. All authors contributed to the writing of the paper.

- Xabier Saralegi and Iñaki San Vicente. Elhuyar at tweetnorm 2013. In *Proceedings of the TweetNorm Workshop at SEPLN*, 2013b

  Contribution to the paper: Both authors contributed equally to the design of the system and the algorithm implemented. First author implemented the code for generating the candidates and carried out the experiments for correct candidate selections. Iñaki San Vicente took care of pre-processing the data, identifying normalization patterns and preparing language models.

CHAPTER 5

Language identification

TweetLID: A Benchmark for Tweet Language Identification

Arkaitz Zubiaga[1], Iñaki San Vicente[2], Pablo Gamallo[3], José Ramom Pichel[4], Iñaki Alegria[5], Nora Aranberri[5], Aitzol Ezeiza[5], Víctor Fresno[6]

[1] University of Warwick, [2] Elhuyar, [3] USC
[4] imaxin|software, [5] University of the Basque Country, [6] UNED

Language identification, as the task of determining the language a given text is written in, has progressed substantially in recent decades. However, three main issues remain still unresolved: (i) distinction of similar languages, (ii) detection of multilingualism in a single document, and (iii) identifying the language of short texts. In this paper, we describe our work on the development of a benchmark to encourage further research in these three directions, set forth an evaluation framework suitable for the task, and make a dataset of annotated tweets publicly available for research purposes. We also describe the shared task we organized to validate and assess the evaluation framework and dataset with systems submitted by seven different participants, and analyze the performance of these systems. The evaluation of the results submitted by the participants of the shared task helped us shed some light on the shortcomings of state-of-the-art language identification

systems, and gives insight into the extent to which the brevity, multilingualism, and language similarity found in texts exacerbate the performance of language identifiers. Our dataset with nearly 35,000 tweets and the evaluation framework provide researchers and practitioners with suitable resources to further study the aforementioned issues on language identification within a common setting that enables to compare results with one another.

Published in *Language Resources and Evaluation*, 50(4):729–766, Dec 2016. ISSN 1574-0218. doi: 10. 1007/s10579-015-9317-4

5.1 Introduction

Recent research shows that while Twitter's predominant language was English in its early days, the global growth and adoption of the social media platform in recent years has increased the diversity in the use of languages (Lehman, 2014). This has in turn fostered an increasing interest of the scientific community in automatically guessing the languages of tweets (Carter et al., 2013). The identification of the language of a tweet is crucial for the subsequent application of widely used NLP tools such as machine translation (Jehl et al., 2012), sentiment analysis (Agarwal et al., 2011; Kouloumpis et al., 2011), Named Entity Recognition (NER) (Li et al., 2012), entity linking (Guo et al., 2013; Cassidy et al., 2012), text summarization (O'Connor et al., 2010; Zubiaga et al., 2012), and lexical (Alegria et al., 2014) and syntactic normalization (Kaufmann and Kalita, 2010), among others. The main problem lies in that this kind of NLP tools tend to be crafted with resources specifically trained for a language or some languages. Hence, these tools cannot deal with unknown languages unless suitable resources are developed. This makes language identification a crucial task especially in multilingual environments such as Twitter, where accurately identifying the language of a tweet enables the application of NLP resources suitable to the language in question.

Twitter itself does provide a language id along with each tweet's metadata, but as we show in this article it leaves much to be desired in terms of accuracy. Besides, it is intended to detect major languages, and does not identify other languages with lesser presence on the platform such as Catalan, Basque or Galician, which account for millions of native speakers within the Iberian Peninsula. In this work, we set out to study the development of language identification systems that deal with more complex situations, including the aforementioned shortcomings of Twitter.

To that end, we first review the related work on language identification and the issues that remain unresolved as of today. Then, we introduce a benchmark dataset and evaluation framework that enables to evaluate different language identification systems, dealing with three of the most important issues that are not resolved: (i) distinction of similar languages, (ii) detection of multilingualism in a single document, and (iii) identifying the language of short texts.

To develop and validate such a benchmark dataset and evaluation framework, we have organized a shared task on tweet language identification (TweetLID), and invited researchers to submit their language identification systems. The task focused on the five most spoken languages of the Iberian Peninsula (Spanish, Portuguese, Catalan, Basque and Galician), and English. These languages are likely to co-occur along with many news and events relevant to the Iberian Peninsula, and thus an accurate identification of the language is key to make sure that we use the appropriate resources for the linguistic processing. This task has intended to bring together contributions from researchers and practitioners in the field, to develop and compare tweet language identification systems designed for the aforementioned languages, which can potentially later be extended to a wider variety of languages. The task meets the aforementioned unresolved issues, given that (i) the task includes four Romance languages which are somewhat similar to one another, (ii) tweets can often be multilingual, and (iii) tweets are short by nature.

This research aims to satisfy the lack of both a benchmark dataset and an evaluation framework to compare different language identification systems. This dataset can be further used by interested researchers and practitioners to make progress in the development of tweet language identification systems.

In this paper, we introduce the benchmark dataset and evaluation framework that enabled the organization of the shared task, which is also made publicly available for research purposes. Then, we analyze and discuss the performance of the different participants of the shared task, which brings to light the most challenging aspects encountered by the participants and need to be addressed in future work. We end by discussing the main objectives that language identification for short texts should pursue in the next years.

This paper substantially extends the overview article we published with the proceedings of the TweetLID workshop (Zubiaga et al., 2014). In this extended paper, we provide an extensive review of the literature, and perform a detailed analysis of the results, by looking among others at numerous aspects relevant to the task, including the three unresolved issues, namely the brevity of texts, multilingualism, and similar

languages. Moreover, this paper discusses the achievements and limitations of the presented systems, summarizing the challenges that are still open for future work.

5.2 Language Identification

Language identification consists in determining the language a text is written in. It has usually been tackled as a classification problem in previous research, often assuming that a document is entirely written in a single language. The best known approaches make use of n-grams to learn the model for each of the languages, as well as to represent each of the documents to be categorized into one of the languages (Cavnar et al., 1994). A language identification system is usually defined as a text classification task (Sebastiani, 2002).

Here we focus on language identification for short texts, more specifically tweets, which is still in its infancy as a research field. Tweets present different characteristics that make the language identification task more challenging. These include that:

- The brevity of the tweets implies that there is very little content that helps to determine the language being used.
- The system allows to use different features along with the content, which do not usually reflect the language of the text. These features include user mentions, hashtags, or retweets, among others.
- Users tend to shorten and/or encode many words in the form of chatspeak, while also introducing typos and misspellings, which deviates the text from its standard spelling.

Provided the aforementioned characteristics inherent in tweets, the language identification for these short texts involves a number of extra challenges that were not considered in other language identification tasks for standard documents such as news stories, books, or even the Web.

5.3 Related Work

In this section, we review previous work in the literature. We start with the historical background of the research in the field of language identification. Then, we summarize

the findings of several comparative studies, and continue by discussing the different directions that research in this field has taken, including language identification for web pages, word level language identification, and language identification for short texts and tweets. We then discuss recent shared tasks that were related to the objectives of TweetLID, and conclude the section by enumerating and discussing the state-of-the-art of the main challenges that our work deals with.

5.3.1 Historical Background

Language identification has attracted a substantial interest in the scientific community in recent decades. While the task was first studied within the community of translators (Beesley, 1988; Newman, 1987; Keesan, 1987; Ingle, 1980) mostly in the 1980s, it started to be more widely studied within the machine learning and natural language processing communities in the 1990s (Cavnar et al., 1994; Dunning, 1994).

Early work on language identification from texts relied on manually defining rules that could be useful in the development of computational tools. For instance, Beesley (1988) proposed relying on language-specific characters to distinguish certain languages, such as ñ or ü for Spanish, or ã for Portuguese. Beesley suggested that such an approach could perform reasonably well for certain languages. However, this approach could perform well for reasonably long and correctly spelled texts in a small set of languages, but more sophisticated techniques might be needed in other scenarios. Later, Cavnar et al. (1994) introduced one of the earliest and most frequently used approaches to language identification in texts: TextCat. Their system computes the n-grams from an input text, and compares the n-grams to the models learned for each of the target languages. The system computes the distance measures with respect to each target language, to assign the language with the lowest distance. This approach achieved 99.8% correct classification rate on Usenet newsgroup articles. Dunning (1994) developed a language identification system using Markov models and a Bayesian classifier. The classifier looks for sequences of characters and words that are unique for each language in the training set, to find similar patterns in the test set. He showed that with only 50k characters of training data, the system could achieve up to 92% accuracy values when identifying the language for short texts of 20 characters. The accuracy increased to more than 99% with larger training sets and test strings with more than 100 characters. He pointed out five key conditions that determine the performance of a language identification system: (i) how the test strings are picked, (ii) the amount of training material

available, (iii) the size of the strings to be identified, (iv) the number of languages to be identified, and (v) whether there is a correlation between domain and language.

In another early attempt, Prager (1999) introduced Linguini, a language identification system which uses n-grams and words as features. The system achieved high performance for classification of monolingual documents in 20 different languages, but its performance dropped significantly for short texts. The author also discussed the applicability of the method to bilingual and trilingual documents. Among the different features studied, 4-grams showed to be the best length for n-grams, and words of unrestricted length did better than considering only short words. The combination of both, 4-grams and words of unrestricted length, performed best. More recently, Lui and Baldwin (2011) developed a method suited to cross-domain language identification. It relies on information gain to identify the features that are strongly predictive of language across domains. Building a feature set from 50,000 documents in 97 languages across 5 datasets, the authors showed that the proposed method can outperform well-known systems such as TextCat (Cavnar et al., 1994) when applied to different domains. Finally, Lui and Baldwin (2012) released langid.py, an off-the-shelf language identification script developed in Python. The script is developed using a Naive Bayes algorithm that relies on n-grams extracted from texts to identify the language, and is intended to be easy-to-use and applicable to different domains.

5.3.2 Comparison Studies

As research in language identification systems made progress, some researchers also conducted comparison studies to find the approaches that work best. Grefenstette (1995) compared two language identification approaches. One using character trigrams as features, and the other one using common short words as features. Their experiments on corpora in 10 European languages showed that either of the compared approaches achieves high accuracy for long texts with more than 50 words, but that trigrams are much more robust for shorter texts. Padró and Padró (2004) compared three statistical methods for language identification: Markov Models, Trigram Frequency Vectors, and n-gram text categorization. They used corpora in 6 different languages for their experiments. They found that Markov Models performed best among the three approaches under study. While the size of the training set did not have a huge impact in the system performance when the training set had at least 50,000 words, they found significant differences in performance when the texts to be classified were very short. Baldwin and Lui (2010) describe a

set of experiments comparing different language identification techniques on three web document datasets. Comparing 1-Nearest Neighbors (1-NN), Naive Bayes, and Support Vector Machines (SVM) with different similarity measures. They found that the most consistent model overall is either a simple 1-NN model with cosine similarity, or an SVM with a linear kernel, using a byte bigram or trigram document representation. They posit that the task becomes increasingly challenging as the number of target languages increases, the size of the training data decreases, and the length of the documents is shorter.

5.3.3 Web-Based Approaches

The emergence of the Web, as an information source that gathers a myriad of documents in an endless number of languages, attracted also a community of researchers to studying language identification approaches in this scenario. Kikui (1996) described a language identification system for online documents. The system was implemented using language models, and could deal with 9 language and 11 coding systems from Eastern Asia and Western Europe. Their experiments on 640 online documents led to a level of accuracy over 95%. On another study on language identification for web pages, Martins and Silva (2005) used the system implemented by Cavnar et al. (1994), complemented with heuristics that specifically deal with HTML markup, and a new similarity measure. They used the web page language identification system to build a search engine that only indexes web content in Portuguese. Their system achieved 99% accuracy in distinguishing Portuguese from the rest of the languages. Xafopoulos et al. (2004) used Hidden Markov Models (HMM) to model character sequences in web documents. Their experiments with web documents in 5 European languages, achieving accuracy values of up to 97%. Baykan et al. (2008) studied the feasibility of determining the language of the content of a web page by only looking at its URL, i.e., without having to download its content. They built a classifier based on keywords extracted from URLs, which was tested on a collection of web pages in 5 languages, achieving 90% in terms of F1 measure. Xia et al. (2009) study the suitability of existing language identification techniques to collections including documents written in one of hundreds of languages, which they motivate as being closer to the nature of the Web. Using the ODIN database[1] for the experiments, which includes documents in nearly a thousand languages, they found that well-known language identification techniques achieved performance values as low as 55%. They introduced a new

[1] http://odin.linguistlist.org/

method which uses context within the document, and formulated the task as a coreference resolution problem, achieving higher performance than using existing techniques for collections with a large number of languages and small training data. Similar to ODIN, the work by Ralf Brown (2012; 2013) has focused on expanding the number of languages considered simultaneously (developing a language identification system for over 1,100 languages). Alongside these works, the Crubadan Project, led by Scannell (2007), aimed at building a large corpus for under-resourced languages using the Web as a source. The project led to the creation of a corpus in more than 400 languages, especially intended for the development of linguistic resources for under-resourced languages.

5.3.4 Word Level Strategies

Motivated by the fact that there are many multilingual speakers who often switch between languages within a sentence, in recent years there is also an increasing interest in the study of word level language identification, i.e., determining what language each word of a sentence is written in. Nguyen and Doğruöz (2014) built a dataset from a Turkish-Dutch community of users, where users mix these two languages, occasionally mixing it with English too. By annotating the language of single words, they experimented with Conditional Random Fields (CRF), which they proved effective at nearly 98% accuracy when using the previous and next tokens to add context to each word. Gella et al. (2014) studied word level language identification for 28 languages, where the system does not know a priori which two languages might co-occur in a text. They defined different heuristics, applied to existing language identification tools such as langid.py and linguini. The heuristics include, for instance, assuming that code-mixing is only likely to occur between certain pairs of languages, but not any possible pair. Their system outperformed existing language identification techniques which are not designed to deal with code-mixed texts, but tends to confuse between languages which are linguistically related. King and Abney (2013) described a weakly supervised language identification system which can be trained using monolingual text samples. Using n-grams as the features to represent the texts, they showed that Conditional Random Fields (CRF) with Generalized Expectation (GE) (Druck, 2011) criteria performed best. The major issue they encountered in the word level identification task were the Named Entities (NE) mentioned in the text, which are very difficult to identify when the language is unknown a priori. They conclude suggesting that a word level language identification system could be built in two steps, the first step

being the high level identification of languages used in a text, and the second step being the specific assignment of language labels to words.

5.3.5 Tweets/Short Messages

Little work has been done on language identification of short texts. Research in this direction has increased especially in recent years, with the advent of social media and microblogs. Tromp and Pechenizkiy (2011) proposed a graph-based n-gram approach for tweet language identification. Using Twitter datasets with monolingual tweets in six languages, they achieved performances between 95% and 98%. Vogel and Tresnerkirsch (2012) extend the work by Tromp and Pechenizkiy by proposing several linguistically-motivated modifications to their algorithm and achieving 99.8% accuracy.

Laboreiro et al. (2013) used a Bayesian classifier to distinguish between European and Brazilian variants of tweets written in Portuguese language, achieving 95% accuracy. Winkelmolen and Mascardi (2011) also describe a Bayesian classifier that performs well on very short texts and made experiments on film subtitles in 22 languages. The work by Murthy and Kumar (2006) deal with short texts, and are especially interested in satisfying the scarcity of research in language identification for a variety of Indian languages, including Hindi, Bengali, Marathi, Punjabi, Oriya, Telugu, Tamil, Malayalam and Kannada. Bergsma et al. (2012) studied different language identification techniques on Twitter datasets with tweets in 9 languages which use Cyrillic, Arabic, and Devanagari scripts. Multilingual tweets were annotated with the predominant language in the tweet, and hence multilingualism was not considered. Given that the dataset includes 3 languages in each of the alphabets, they divide the task into 3 smaller subtasks. They tested three language identification systems, using textual features such as n-grams, and user metadata from Twitter, as well as Wikipedia as an external resource. They showed that by combining n-grams and user metadata, their system can achieve up to 98% accuracy in each subtask that deals with three languages. Goldszmidt et al. (2013) tested statistical language identifiers, based on character frequencies, to classify tweets in five different languages by using Wikipedia for training. While they found that Wikipedia is insufficient to represent several idioms used exclusively in social media, they introduced a boot-strapping technique that significantly improves the accuracy of the language identifier. Hammarstrom (2007) described a fine-grained model which stores a large frequency dictionary as well as an affix table and is able to classify with high accuracy short texts of just one word.

Carter et al. (2013) investigated language identification on a Twitter dataset with tweets in five major European languages: Dutch, English, French, German, and Spanish. To enrich the textual content of tweets, they use additional context surrounding the tweets: (i) the content of the link being pointed to, (ii) the author of the tweet, (iii) mentions of other users, (iv) context from the tweet that it is replying to, and (v) hashtags. They found the combination of all five features to perform best. In our work, we argue that the collection of such context for each tweet is time-consuming, and makes it impossible to run the language identifier in a timely fashion for a relatively large set of tweets. To account for this, we present a tweet dataset and describe the problem as a task where the language of a tweet has to be determined from its readily available features.

Lui and Baldwin (2014) presented an evaluation of several language identification systems applied to tweets. They showed that simple voting over three specific systems consistently outperforms any specific system, and achieves state-of-the-art accuracy on the task. In addition, the authors also defined a semi-automatic method to construct annotated datasets of tweets for evaluating a language identification system.

In a comparative study where a number of well-known language identification systems were tested on a Twitter dataset with tweets in five languages, Derczynski et al. (2015) showed that Cavnar and Trenkle's TextCat (1994), retraining its models based on tweets, performed best. This comparison also shows a big difference between training TextCat in tweets (97.4% accuracy), or using its own models (89.5% accuracy).

5.3.6 Related Shared Tasks

In recent years, there have been several shared tasks on language identification, which are relevant to the shared task we organized at TweetLID. The 2010 Australasian Language Technology (ALTA-2010) organized a workshop and shared task on Multilingual LangID. The dataset for the task was created by Baldwin and Lui (2010) from editions of Wikipedia in different languages. In 2013, the workshop on Innovative Use of NLP for Building Educational Applications (BEA8)[2], co-located with NAACL, hosted a shared task on Native Language Identification (NLI). The task consisted in identifying the native language of a writer based solely on a sample of their writing (Tetreault et al., 2013). Another relevant shared task is Language

[2]http://www.cs.rochester.edu/ tetreaul/naacl-bea8.html

Identification in Code-Switched (CS)[3], which was part of the First Workshop on Computational Approaches to Code Switching, organized within the EMNLP-2014 conference. This shared task focused on short texts having in than one language. Moreover, the shared task Discriminating Similar Languages (DSL-2014)[4], organized within COLING-2014, deals with discriminating between similar languages and language varieties, which is one of the bottlenecks of language identification.

5.3.7 Challenges

Among the little work on the study of language identification techniques for tweets, no research has dealt so far with code-mixing and the identification of multilingualism in tweets, and no special attention has been paid to similar languages in these short texts. Our work looks specifically at these two aspects, multilingualism and similar languages, in the context of short texts.

Others have looked at additional challenges that can occasionally be also part of a language identification task. Chepovskiy et al. (2012) looked at how to deal with language identification of transliterated texts. They explored the ability to identify five Slavic languages from their Latin transliterations. Also, Sibun and Spitz (1994) studied the accuracy of language identification systems when applied to scanned images. Our work, instead, assumes that the input sentences are given as texts.

Regarding the identification of similar languages, Ljubešić et al. (2007) studied the case of Croatian, which language identification tools find it hard to distinguish from similar languages such as Serbian, Slovenian, or Slovak. By defining a set of rules that specifically characterize the Croatian language, such as identifying the most frequent words, their system outperformed existing tools.

Language identification has progressed significantly in recent years, to the point that the task has been considered solved for certain situations (McNamee, 2005), assuming among others that documents are long enough and that are written in a single language. However, the emergence of social media and the chatspeak employed by its users has brought about new previously unseen issues that need to be studied in order to deal with these kinds of texts. Three key issues posited in the literature (Sibun and Reynar, 1996; Hughes et al., 2006; Řehůřek and Kolkus, 2009) and that, as of today, cannot be considered solved include: (i) distinguishing similar languages (Zampieri, 2013), (ii) dealing with multilingual documents (Lui and Baldwin, 2014),

[3] http://emnlp2014.org/workshops/CodeSwitch/call.html

[4] http://corporavm.uni\discretionary{-}{}{}koeln.de/vardial/sharedtask.html

and (iii) language identification for short texts (Bergsma et al., 2012; Carter et al., 2013; Laboreiro et al., 2013; Gottron and Lipka, 2010; Vatanen et al., 2010; Nguyen and Doğruöz, 2014). The shared task organized at TweetLID has considered these three unresolved issues, and has enabled participants to compare the performance of their systems in these situations.

5.4 Defining the Tweet Language Identification Task

Within the linguistically diverse nature of social media, and specifically Twitter in our case, we set forth the tweet language identification task as the problem that consists in identifying the language or languages tweets are written in. In this work, we have created a Twitter dataset that enables to study language identifiers in a context where tweets are of multilingual nature, often due to the users' tendency to code-mixing, and there is a high degree of similarity between some of the languages. This dataset has been tested in a shared task, TweetLID (Zubiaga et al., 2014), which allowed participants to evaluate their language identifiers in a common setting. The dataset and task focused on the most widely used languages of the Iberian Peninsula, which provides an ideal context where news and events are likely to be shared and discussed in multiple languages.

To the end of setting up a common evaluation framework to enable comparison of different language identification systems, we put together an annotated corpus of nearly 35,000 tweets and defined a methodology to evaluate the multi-label output of the language identification systems. Splitting the corpus into a training set with 15k tweets, and a test set with 20k tweets, the participants had a month to develop their language identification systems making use of the training set. They then had 72 hours to work on the test set and submit their results. The shared task consisted of two separate tracks: (1) constrained, where external corpora could not be used for training, and (2) unconstrained, where the use of external corpora was permitted. Each participant could participate with up to two submissions per track.

Besides the challenge of dealing with the short and often informal texts found in tweets, the task considered that a tweet is not necessarily written in a single language. This is especially true in bilingual regions, where speakers that feel equally comfortable with either of their two native languages tend to code-switch between them and mix them in a sentence quite frequently (Cárdenas-Claros and Isharyanti, 2009; Myers-Scotton, 2002; Paolillo, 2011). Hence, the task also considered a number

of cases where the response is not basically one of the languages in the list: (i) a tweet can combine two –or occasionally three– languages in a tweet, e.g., when a tweet has parts in Catalan and Spanish, (ii) given the similarity and cultural proximity between some of the languages, it is not possible to determine which of two –or more– languages a tweet is written in, e.g., some tweets might be written equally in Catalan or Spanish, (iii) despite the geographical restriction of the tweets in the task, it is also likely that tweets in other languages occur, such as French, and (iv) it is not possible to determine which of the 6 languages considered in the task a tweet is written in, e.g., when a tweet only mentions entities, smileys, or onomatopoeias. We will elaborate more on these cases in the next section introducing the dataset and the annotation process.

The dataset includes the five top languages of the Iberian Peninsula, which are spoken in different regions, and four of them –Spanish, Portuguese, Catalan, and Galician– are romance languages originating from Latin and with certain similarities among them, which makes the task more challenging. The fifth language –Basque–, and English, belong to different language families, and therefore are rather different from the rest. Still, their cultural proximity, and the fact that many users in the area are bilingual, entails that they often mix words and spellings across languages. For instance, a Basque native might naturally write something like *"nos vemos, agur!"* (see you later, bye!), when *"nos vemos"* is in Spanish, and *"agur"* is Basque to say good bye; similarly, a Catalan speaker might often misspell the Spanish word *"prueba"* (test) as *"prueva"*, given that the Catalan translation of the word (*"prova"*) is written with *v*. These characteristics are common in bilingual areas, and have been considered in the definition of this task in order to carefully develop the annotation guidelines and to pursue the final annotation of the corpora.

5.5 Creation of a Benchmark Dataset and Evaluation Framework

In this section, we first describe the process we followed to collect data from Twitter, then we explain how we annotated manually the tweets with the language label in question, and finally we describe the evaluation measures we used for the task.

5.5.1 Data Collection

To collect an unrestricted set of tweets, but rather focused on the set of languages within the scope of TweetLID, we relied on geolocation to retrieve tweets posted from areas of interest. We used Twitter's streaming API's `statuses/filter` endpoint to collect geolocated tweets posted within the Iberian Peninsula from March 1 to 31, 2014. While this stream is limited to tweets explicitly providing geolocation metadata, it allows to track a diverse set of tweets that is not restricted to a specific set of users or domain. Having collected these tweets, we used Nominatim[5] to obtain specific location information for each tweet. Given the coordinates of a tweet as input, Nominatim queries OpenStreetMap for the specific address associated with those coordinates, i.e., region, city, and street (if available) from which the tweet has been sent. This led to the collection of 9.7 million tweets with location details associated. From this set of tweets, we sampled tweets from **Portugal** and the following **3 officially bilingual regions**:

- **Basque Country**, where Basque and Spanish are spoken. Tweets from the province of Gipuzkoa were chosen here to represent the Basque Country.
- **Catalonia**, where Catalan and Spanish are spoken. Tweets from the province of Girona were chosen to represent Catalonia.
- **Galicia**, where Galician and Spanish are spoken. Tweets from the province of Lugo were chosen.

One province was picked from each of the regions to avoid cases such as that of the province of Barcelona in Catalonia, which is much more diverse in terms of languages due to tourism. These three bilingual regions enabled us to sample tweets in Basque, Catalan, Galician, and Spanish, and we could sample Portuguese tweets from Portugal. English is the sixth language in the corpus, which can be found all across the aforementioned regions. For the final corpus to be manually annotated, we picked 10k tweets from each of the bilingual regions, and 5k from Portugal. The tweets picked here had to contain at least one word (i.e., string fully made of a-z characters), so that there is some text, and tweets with e.g. only a link are not considered. The next section describes the manual annotation performed on this corpus with 35k tweets.

[5]http://wiki.openstreetmap.org/wiki/Nominatim

5.5.2 Manual Annotation

The collection of 35k tweets resulting from the aforementioned process was then manually annotated. Each of the tweets was associated with its corresponding language code in the manual annotation process. The manual annotation was conducted by annotators who were native or proficient speakers in at least three languages considered in the task. This enabled us to distribute the tweets from each of the four regions to different annotators, so that each annotator was a native or proficient speaker of the languages spoken in the region in question, as well as English.

The annotators were instructed to assign codes to tweets according to the language in which they were written. We asked them to ignore #hashtags and @user mentions, as well as references to NEs in another language. For instance, in the tweet *Acabo de ver el último capítulo de la temporada de 'the walking dead', muy bueno!* (Spanish: I just saw the season finale of 'the walking dead', it's amazing!), only Spanish should be annotated, irrespective of the named entity 'the walking dead' being in English.

They had to assign codes to the tweets as follows: *eu* for Basque, *ca* for Catalan, *gl* for Galician, *es* for Spanish, *pt* for Portuguese, and *en* for English. When a different language was found in a tweet –e.g., French or German–, they had to annotate it as *other*. Additionally, when the text of a tweet included words that are widely used in any of the languages in the task –e.g., onomatopoeias such as 'jajaja' or 'hahaha', or internationalized words such as 'ok'–, which makes it impossible to determine the language being used in that specific case, they were asked to annotate it as *und*(eterminable). These eight cases —i.e., eu, ca, gl, es, pt, en, other, und– constitute all the options for **monolingual tweets**.

In the above situations, the annotators had to mark a tweet as either being written in one of the 6 languages, *other* or *und*. However, two more cases were identified and included in the annotation guidelines: ambiguous tweets, and multilingual tweets.

Ambiguous tweets were defined as those that can be categorized into the list of languages being considered, but may have been written in at least two of them. Given the similarity and cultural proximity of some of the languages, it is likely that some short texts are written equally in some languages. For instance, *Acabo de publicar una foto* (I just published a photo) can be either Spanish or Catalan, and cannot be disambiguated in the absence of more context. This case had to be annotated as *es/ca*.

Multilingual tweets contain parts of a tweet in different languages, where the annotators were instructed to annotate all of the languages being used. For instance, *Qeeeee matadaaa* (Spanish: that was exhausting) *da Biyar laneaaaa...* (Basque: and gotta go to work tomorrow) should be annotated as *es+eu*, and *Acho que vi a Ramona hoje* (Portuguese: man, I've seen Ramona today) *but im not sure* (English) should be annotated as *pt+en*. Occasionally, three languages were also found, e.g., *Egun on! Buenos días! Good morning!* (Good morning in Basque, Spanish and English), annotated as *eu+es+en*. The annotation had to consider all the languages being used, in no specific order, except when a single word or term was used as a constituent of a sentence in another language, e.g., *es un outsider* (Spanish: he is an outsider), where only one language is annotated.

The last possible cases are the **mixed tweets**, which are the result of having multilingual tweets where at least one of the languages is either *undeterminable*, *other*, or *ambiguous*. It could also be the case that a multilingual tweet with two languages is the combination two of the cases above, e.g., *other + ambiguous*. However, we have not found any of these cases in our dataset. We have only found cases where one of the six languages under study is combined with either *other* or *ambiguous*, which were ultimately removed from the dataset for being very rare and not having enough examples for training, as we describe next.

5.5.3 Annotated Corpus and Evaluation Measures

All the 35,000 tweets were annotated following the aforementioned methodology. Given that the cases where a tweet was annotated as a *mixed tweet* –i.e., where certain language was combined with a language not considered in the task ('lang+other'), or with an ambiguous text ('lang1+lang2/lang3')– were very rare, they were removed from the dataset. These include only 16 cases, which after removing led to an annotated corpus composed of 34,984 tweets. The corpus, including also the content of the tweets, can be found on the shared task's web site[6]. Table 5.1 shows the distribution of the manual annotations, where it can be seen that Spanish is the predominant language, which amounts to 61.22% of the tweets. This is why we use a macroaverage approach to evaluate the systems, as we describe later, which rewards the systems that perform well for all the languages rather than just for the predominant language. Table 5.2 shows a breakdown of the annotations by region. It shows that the prevalence of Spanish is especially marked in Galicia (86.61%) and in the Basque Country (78.46%). It is more evenly distributed in Catalonia, with

[6]http://komunitatea.elhuyar.org/tweetlid/resources/

50.62% of the tweets in Spanish and 29.40% in Catalan. Spanish barely occurs in Portugal (only 1.16% of the times), where Portuguese is the predominant language with 81.82% of the tweets. English has a moderate presence across all regions, ranging from 1.55% to 8.28%, and the other three languages –Catalan, Basque, and Galician– have a tiny presence outside their region. The number of ambiguous tweets is much higher in Portugal than in the other regions, especially due to the large number of Portuguese tweets that could also be deemed Galician (pt/gl). Multilingual tweets occur especially in the Basque Country (mostly eu+es), given that code switching occurs very often in this region, and the fact that the two languages are so different makes it easy for the human annotator to identify the presence of both languages; likewise, due to the big difference between both languages, it is very unlikely that a tweet is ambiguous in Spanish or Basque (eu/es). The number of "other" languages is significantly higher in Catalonia than in the other regions, potentially due to the higher diversity of nationalities, due to being a rather touristic region, and a close-by region for the French and Italians, whose languages are considered as "other" in this work.

Language	**Tweets**	**%**
Spanish (es)	21,417	61.22
Portuguese (pt)	4,320	12.35
Catalan (ca)	2,959	8.46
English (en)	1,970	5.63
Galician (gl)	963	2.75
Basque (eu)	754	2.16
Undeterm. (und)	787	2.25
Multilingual (a+b)	747	2.14
Ambiguous (a/b)	625	1.79
Other	442	1.26

Table 5.1: Distribution of the manual annotation.

Additionally, we asked a second annotator for each of the regions to re-annotate a 10% sample of the tweets, i.e., 3,500 tweets altogether. This allows us to compute the inter-annotator agreement on a 10% sample of the whole, so that we can measure the difficulty of the task for human annotators. The inter-annotator agreement is computed as the pairwise agreement between the two annotations for each tweet. Only exact matches are considered as agreement, hence if an annotator labeled

	Basque Country		Catalonia		Galicia		Portugal	
Language	**Tweets**	**%**	**Tweets**	**%**	**Tweets**	**%**	**Tweets**	**%**
Spanish (es)	7842	78.46	5057	50.62	8460	84.61	58	1.16
Portuguese (pt)	22	0.22	44	0.44	163	1.63	4091	81.82
Catalan (ca)	20	0.20	2937	29.40	1	0.01	1	0.02
English (en)	595	5.95	827	8.28	155	1.55	393	7.86
Galician (gl)	2	0.02	2	0.02	959	9.59	0	0.00
Basque (eu)	751	7.51	2	0.02	1	0.01	0	0.00
Undeterm. (und)	233	2.33	386	3.86	34	0.34	134	2.68
Multilingual (a+b)	430	4.30	230	2.30	40	0.40	47	0.94
Ambiguous (a/b)	65	0.65	137	1.37	167	1.67	256	5.12
Other	35	0.35	368	3.68	19	0.19	20	0.40

Table 5.2: Distribution of the manual annotation by region.

a tweet as "gl", and the other annotated it as "es/gl", this is computed as a disagreement. Overall, the annotators agreed 92.6% of the times, distributed by region as shown in Table 5.3. These values show that, to some extent, the distinction between similar languages as well as very frequent linguistic interferences can make it difficult for the human annotator. This can be observed especially in the case of Galicia, where the inter-annotator agreement rate is lower than for the other regions. The low inter-annotator agreement values between "es" and "gl" in Galicia can be explained by two factors: on the one hand, the official Galician language uses the same spelling system as Spanish and, on the other hand, the colloquial Galician language often contains many Spanish interferences since people tend to make use of informal Spanish words and expressions. This makes the distinction between the two languages an even more challenging task for human annotators. It is also worth mentioning that while the annotation work for each region will mostly include tweets involving the two languages spoken in the region, there are multiple combinations of those (e.g., es, gl, es+gl, es/gl), besides the fact that other languages also occasionally occur.

Moreover, we also wanted to look at two more factors that are key in our research goals: (1) the length of tweets, to check whether the brevity also makes it more difficult for human annotators, and (2) the fact that tweets are monolingual or multilingual. Table 5.4 shows the agreement values broken down by length. The agreement rates show that there is no significant difference for tweet lengths ranging from 21 to 140 characters. However, the agreement rate drops for tweets between 1 and 20 characters; a number of these cases where due to the difficulty of distinguishing

Region	Agreement	Most frequent errors
Basque Country	93.6%	es → en+es (1.20%) es+eu → es (0.50%) es+eu → eu (1.00%) en+es → es (0.50%)
Galicia	88.1%	gl → es (4.20%) en → es (1.90%) und → es (1.90%) es/gl → es (1.00%) es → gl (1.00%) es → es/gl (0.50%)
Catalonia	96.0%	es → ca/es (0.50%)
Portugal	93.0%	pt → gl/pt (1.20%) gl/pt → pt (1.20%)

Table 5.3: Inter-annotator agreement values distributed by region, computed as the pairwise agreement between two annotators for 10% of the corpus. The last column of the table shows the most frequent disagreements between annotators.

whether a short tweet is written in a certain language or is instead undeterminable, while other cases include confusions between Galician and Spanish or Portuguese, as well as English with Spanish, e.g., due to barbarisms. On the other hand, Table 5.5 shows the agreement values for monolingual and multilingual tweets. In this case, the agreement rate is substantially lower for multilingual tweets than it is for monolingual tweets. The errors when annotating multilingual tweets include a majority of cases where an annotator labeled a tweet as being only Spanish, while the other labeled it as being in both Spanish and English; again, this depends on each annotator's judgment on whether an English word in a Spanish sentence is a barbarism, or can be considered as a constituent word in Spanish.

The manual annotation work was performed separately for each region, especially given that this facilitates the annotators' work, and it does not require proficient knowledge of the six languages under study. The shared task, however, puts together all the tweets from the four regions, where the language identifiers need to identify all the languages in the same task.

For the purposes of the shared task, the corpus was split into two random sets of

Tweet length	Agreement	Most frequent errors
121-140	92.5%	es → es/gl (1.49%) es → es+gl (1.49%) es → pt (1.49%) en → en+es (1.49%) en → es (1.49%)
101-120	94.4%	es → ca (1.85%) en → es (1.85%) gl → es (1.85%)
81-100	94.5%	en → es (2.06%)
61-80	96.2%	gl → es (3.08%)
41-60	94.8%	gl → es (1.48%) es → en+es (0.74%) und → es (0.74%)
21-40	91.9%	gl → es (2.28%) es/gl → es (0.98%) und → es (0.98%) en → es (0.65%)
1-20	82.5%	und → es (2.92%) es → es/gl (2.19%) und → en (1.46%) en → es (1.46%) pt → gl/pt (1.46%)

Table 5.4: Inter-annotator agreement values by tweet length.

Monolingual/multilingual	Agreement	Most frequent errors
Monolingual	94.3%	gl → es (1.79%) und → es (1.05%) en → es (0.95%) es/gl → es (0.42%) es → gl (0.21%) und → en (0.21%) und → pt (0.21%)
Multilingual	64.7%	es → en+es (11.76%)

Table 5.5: Inter-annotator agreement values for monolingual and multilingual tweets.

tweets: a training set with 14,991 tweets, and a test set with 19,993 tweets. However, due to restrictions on the use of the Twitter API[7], we distributed the corpora to the participants by including only the tweet IDs. We also provided them with a script to download the content of the tweets having the IDs, which scrapes the web page of each tweet to retrieve the content.

Once the participation period ended we checked the set of tweets in the test set that were still available at the moment. This was done specifically on the 7th of July, with the submission deadlines closed for all the participants. This final check found that 18,423 out of the initial 19,993 tweets, i.e., 92.1%, were available at the moment. For further details into the composition of the corpora, Table 5.6 shows the distribution of categories for the train and test datasets.

While the reduction of the evaluation dataset to the 92.1% subset was inevitable at the time the shared task took place, the most recent Terms of Service introduced by Twitter allow us to release the content of the tweets along with the dataset. The fact that new tweets may continue to disappear from Twitter's API does no longer affect to the entirety of the dataset then, and will enable additional research using the original dataset. In order to be able to compare results with those of the participants of the task, we also release the list of 18,423 tweet IDs we used for evaluation.

The participants had to submit their results formatted as 'tweet' and 'lang' pairs, referring to the language each tweet in the test set is written. To be considered a valid response, 'lang' can take one of the following forms:

- 'lang1': single language. Possible values are: [es, en, gl, ca, eu, pt, und, other]

[7]https://dev.twitter.com/terms/api-terms

- 'lang1+lang2[+lang3]': multiple languages. Any combination of the aforementioned codes are allowed.

It is important to note that 'lang1/lang2[/lang3]' was not a valid answer. If such notation was found, only the first language was taken into account.

When using multiple languages, ('lang1+lang2[+lang3]') a maximum number of 3 languages could be included. If in any case more were provided, the first 3 languages will be taken into account.

Language	**%Tweets Train**	**%Tweets Test**
Spanish (es)	57.11 (8,562)	64.02 (11,794)
Portuguese (pt)	14.35 (2,151)	10.55 (1,943)
Catalan (ca)	9.78 (1,466)	7.79 (1,435)
English (en)	6.66 (999)	4.97 (914)
Galician (gl)	3.38 (507)	2.30 (423)
Basque (eu)	2.53 (380)	1.94 (358)
Undeterm. (und)	1.25 (188)	3.01 (555)
Multilingual (a+b)	2.47 (371)	1.93 (356)
Ambiguous (a/b)	2.31 (346)	1.41 (260)
Other	0.14 (21)	2.09 (385)

Table 5.6: Distribution of the manual annotation in train and test data sets.

Evaluation Measures

The fact that the corpora (as well as the reality of Twitter itself) is imbalanced, where some languages are far more popular than others, is an important issue to be considered when defining the evaluation measures. Besides, given that the language identification task has been defined as a classification problem where tweets can be either multilingual, with more than a language per tweet, or ambiguous, where it is not possible to disambiguate among a set of target languages, the evaluation measures need to be carefully defined to take these into account.

To deal with the imbalance, we compute the precision, recall, and F1 values for each language, and the macroaveraged measures for all languages afterwards. This is intended to provide higher scores to systems that perform well for many languages,

rather than those performing very well in the most popular languages such as Spanish or Portuguese. We compute Precision (P), Recall (F) and F1 measures as defined in Equations 5.1, 5.2, and 5.3.

$$P = \frac{1}{|C|} \sum_{i \in C} \frac{TP_i}{TP_i + FP_i} \tag{5.1}$$

$$R = \frac{1}{|C|} \sum_{i \in C} \frac{TP_i}{TP_i + FN_i} \tag{5.2}$$

$$F_1 = \frac{1}{|C|} \sum_{i \in C} \frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i} \tag{5.3}$$

where $C = \{ca, en, es, eu, gl, pt, amb, und\}$ is the set of labels defined in our classification task, and TP, FP and FN refer to the counts of true positive, false positive and false negative answers respectively.

The evaluation of our task needs to deal with a ground truth which is occasionally multi-label, so that traditional approaches used in language identification tasks for computing TP, FP and FN are not directly applicable. For this purpose, we adapt a concept-based evaluation methodology for multi-label classification (Nowak et al., 2010) to the specific purposes of the task, which we further describe next. To determine whether a system's output for a tweet is correct, we compare it with the manually annotated ground truth. Given that tweets are not simply multilingual, the TP, FP and FN values are computed as follows:

- For monolingual tweets, the TP count is incremented by 1 if the answer is correct, and FP is incremented by 1 for the language output by the system otherwise. If a system's prediction contains more than one language, incorrect languages will be penalized, e.g., for a tweet annotated as "pt", a system that outputs "pt+en" will increment TP for "pt" but also FP for "en". FN will be incremented for the language in the ground truth if the answer does not contain the correct language. Hence, the system that outputs "eu" for a tweet that is actually "pt", will count as an additional FP for "eu", and as a FN for "pt".

- For multilingual tweets, we apply the same evaluation methodology as for the multilingual tweets above repeatedly for each of the languages in the ground truth, e.g., for a tweet annotated manually as "ca+es", a system that outputs just "ca" will count as TP for "ca" and as FN for "es".

- For ambiguous tweets that could have been written in any of a set of languages, any of the responses in the ground truth is deemed correct, e.g., for a tweet annotated as "ca/es", either "ca" or "es" is deemed correct as a response, counting as TP of the "amb" category in either case. If, instead, the system outputs "pt", which is not among the languages listed in the ground truth of the ambiguous tweet, the evaluation counts as a FP for "pt", and as a FN for "amb".

Finally, note that we merged tweets annotated as "other" or "und" for evaluation purposes. We did not differentiate between them as those are the tweets that need to be ruled out for being out of the scope of the task. If a system determines that a tweet is "other", and the ground truth is "und", or vice versa, it is deemed correct. To facilitate replication of the experiments as well as comparison of performance results, the evaluation script we used to compute the performance scores is also available on the workshop site.

5.6 Shared Task to Test and Validate the Benchmark

The TweetLID shared task consisted of two separate tracks, one being constrained where external resources were not allowed, and the other being unconstrained were the participants could make use of external resources. Out of the initially registered 16 participants, 7 groups submitted their results for either one or both of the tracks. Participants had a 72 hour window to work with the test set and submit up to two results per track. Next, we first summarize the types of approaches that the participants relied on, and further detail the technique used by each of the participants afterwards.

5.6.1 Overview of the Techniques and Resources Employed

The participants relied on very diverse and different techniques in their systems. They employed different classification algorithms, different methods to learn the models for each language, as well as different criteria to determine the languages of a tweet. This diversity of approaches enables us to broaden the conclusions drawn from the analysis of the performance of different systems. One aspect that the participants agreed upon is the need to preprocess tweets by removing some tokens that do not help for the language identification task such as URLs and user mentions, as well as by lowercasing and reducing the repetition of characters, among others.

TEAM	Classifier	Representation	Ext. Resources	Multiling.
Citius-imaxin	1) ranked n-grams 2) naive bayes	words & n-grams & suffixes	news corpora	no
RAE	support vector machines	n-grams	-	yes
UB/UPC /URV	1) linear interpolation 2) out-of-place measure	n-grams	-	no
IIT-BHU	n-gram distances	n-grams	-	no
CERPAMID	n-gram distances	3-grams	Europarl corpus Wikipedia	no
ELiRF @ UPV	1) support vector machines 2) Freeling	words & 4-grams	Wikipedia	yes
LYS @ UDC	TextCat & langid.py & langdetect	-	Yali	no

Table 5.7: Summary of the main characteristics of the systems developed by the participants

The participants used different classification algorithms to develop their systems. The classification algorithms used by most participants include Support Vector Machines (SVM), and Naive Bayes, which have proven effective in previous research in language identification for longer texts.

Not all the participants developed multilabel techniques that can deal with multilingual tweets. Only two of them actually did, mostly by defining a threshold that determines the languages to be picked for the output when the classifier provides a higher confidence score for them.

Table 5.7 summarizes the characteristics of the approaches developed by each of the participants.

5.6.2 Brief Description of the Systems

Citius-imaxin (Gamallo et al., 2014) submitted two different systems to each of the tracks. On the one hand, a system they called Quelingua builds dictionaries of words ranked by frequency for each language. New tweets are categorized by weighing the ranked words in it, as well as specific suffixes that characterize each language. On the other hand, they build another system based on Naïve Bayes, which has proven accurate in recent research. For the unconstrained track, they fed the systems with news corpora extracted from online journals for all six languages. Their systems do not pick more than one language per tweet, hence not dealing with multilingual

tweets. Their bayesian system achieved the best performance for the unconstrained track. Moreover, it was the only system in the task that outperformed its constrained counterpart.

RAE (Porta, 2014) submitted two systems only to the constrained track. Their systems rely on n-gram kernels of variable length for each language. The best parameters for each kernel were estimated from the results on the unambiguous examples in the training dataset by cross-validation. They then used Support Vector Machines (SVM) to categorize each new tweet. They relied on a decision tree to interpret the output of the one-vs-all SVM approach, and thus deciding whether the confidence values for more than one language exceeded a threshold (multilingual tweet), only one did (monolingual tweet), or none did (undeterminable).

UB/UPC/URV (Mendizabal et al., 2014) submitted one system to each of the tracks. They developed a different type of system in this case for each track. The first system, submitted to the constrained track, makes use of a linear interpolation smoothing method (Jelinek, 1997) to compute the probabilities of each n-gram to belong to a language, and weigh new tweets using those probabilities. The second system, submitted to the unconstrained track, is an out-of-place approach that builds a ranked list of n-grams for each language in the training phase, and compares each new tweet with these ranked lists to find the language that resembles in terms of n-gram ranks.

IIT-BHU (Singh and Goyal, 2014) only submitted a run to the constrained track. They adapted a system that they previously created for other kinds of texts (Singh, 2006), which is a simple language identification system that makes use of n-grams, and based on that created by Cavnar and Trenkle (Cavnar et al., 1994), to the context of Twitter. Basically, they integrated a preprocessing module that removes noisy tokens such as user mentions, hashtags, URLs, etc., and then uses a symmetric cross entropy to measure the similarity or distance between each new tweet and the models learned for each language in the training phase.

CERPAMID (Zamora et al., 2014) submitted two systems to each of the tracks. They extract n-grams of three characters to represent the tweets, and use three different weighing methods to weigh the n-grams. Then, they give a score to each new tweet for all the languages in the collection using the three weighing schemes, and pick the final language given as output by the system through simple majority voting. As their systems only output one language, they did not develop any solutions to deal with multilingual tweets. For the unconstrained track, they used the Europarl corpus (Koehn, 2005) for English, Spanish, and Portuguese, and Wikipedia for Basque, Catalan, and Galician.

ELiRF @ UPV (Hurtado et al., 2014) submitted two systems to each of the tracks. For the constrained track, the authors made use of a one-vs-all classifier combining method using SVM. The two approaches submitted to the constrained track differ in the way they deal with multilingual tweets: on one of the approaches, they consider each combination of languages as a new category, while in the other approach they defined a threshold so that the output included all the languages for which the SVM classifier returned a higher confidence value. For the unconstrained track, they developed a classifier using SVM, which used Wikipedia to train the system but did not return multilabel outputs, and another classifier using Freeling's language identification component (Padró and Stanilovsky, 2012), which includes its own models of 4-grams for the languages in the corpus, except for Basque that the authors created themselves. The constrained method that relies on a threshold to pick the languages for the output achieved the best performance for the constrained track.

LYS @ UDC (Mosquera et al., 2014) submitted two systems to each of the tracks. They used three different classifiers to develop their systems: TextCat (Cavnar et al., 1994), langid.py (Lui and Baldwin, 2012), and langdetect (Shuyo, 2010). The two different systems they developed for both tracks differ in that one determines the final output by relying on the classifier with higher confidence, while the other determines the output by majority voting. For the unconstrained track, they used the corpus provided with Yali (Majliš, 2012). Their systems return a single language as output, not dealing with multilingual tweets.

5.6.3 Results

Table 5.8 shows the results for the *constrained* track, and Table 5.9 shows the results for the *unconstrained* track. The **ELiRF @ UPV** group, with an SVM-based approach that uses 4-grams and words as features, performed best for the constrained track with an F1 of 0.752. In the unconstrained track, **Citius-imaxin** presented the most accurate system with a very similar F1 value, 0.753, which uses a bayesian classifier with words, n-grams and suffixes as features.

One of the aspects that stands out from the results of the participants is the fact that most of the systems performed better in the constrained track, and the lower performance of their unconstrained counterparts suggests that either the external resources used are not suitable for the task, or they were not properly exploited. Surprisingly, the only unconstrained algorithm outperforming its constrained

#	TEAM	P	R	F1
1	ELiRF @ UPV II	0.825	0.744	0.752
2	ELiRF @ UPV I	0.824	0.730	0.745
3	UB/UPC/URV	0.777	0.719	0.736
4	RAE II	0.806	0.689	0.734
5	RAE I	0.811	0.687	0.733
6	Citius-imaxin II	0.824	0.685	0.726
7	Citius-imaxin I	0.689	0.743	0.699
8	CERPAMID I	0.716	0.681	0.666
9	LYS @ UDC I	0.732	0.734	0.638
10	IIT-BHU	0.605	0.670	0.615
11	CERPAMID II	0.704	0.578	0.605
12	LYS @ UDC II	0.610	0.582	0.498

Table 5.8: Performance results for all the submissions to the constrained track, sorted by F1 measure.

counterpart was that by Citius-imaxin. This posits an important caveat of the presented systems, which needs to be further studied in the future.

Next, we delve into the performance of the different systems, by looking at the results broken down into different aspects, which allows us to carry out a more detailed analysis of their performance. First, we perform an alternative microaveraged evaluation of the systems, to complement the analysis. Then, we show the performance of baseline approaches, and compare them with the performance of the participants of the shared task. We then analyze each system's performance in more detail, by looking at the three main issues that motivated our work, i.e., brevity of tweets, multilingualism, and similarity between languages. Finally, we analyze the errors of the systems to better understand the limitations of the language identification systems.

Alternative Microaveraged Evaluation

For the sake of comparison with the performance reported in other research works, we also show here the microaveraged evaluation of the three best systems in each track. Note that the micro-averaged evaluation favors the overall performance of

#	TEAM	P	R	F1
1	Citius-imaxin II	0.802	0.748	0.753
2	ELiRF @ UPV II	0.737	0.723	0.697
3	ELiRF @ UPV I	0.742	0.686	0.684
4	Citius-imaxin I	0.696	0.659	0.655
5	LYS @ UDC I	0.682	0.688	0.581
6	UB/UPC/URV	0.598	0.625	0.578
7	LYS @ UDC II	0.588	0.590	0.571
8	CERPAMID I	0.694	0.461	0.506
9	CERPAMID II	0.583	0.537	0.501

Table 5.9: Performance results for all the submissions to the unconstrained track, sorted by F1 measure.

the systems, regardless of their likely poor performance for some of the languages. Tables 5.10 and 5.11 show the microaveraged results for both tracks, with an overall boost in the results for all the contestants. Still, the best results obtained in this shared task are far from the 99.4% accuracy score reported for formal text, or the 92.4% accuracy score reported for microblogs by Carter et al. (2013). However, it is worth mentioning that Carter et al's scores rely on a monolingual tweet language identification task for major languages including Dutch, English, French, German, and Spanish. The fact that TweetLID has introduced multilingual tweets, as well as tweets from underrepresented languages led to slightly lower performances scores of 89.8% accuracy in the best case. Still, this only reflects a 2.6% accuracy loss when compared to Carter et al's best results for tweets.

Comparison with Baseline Approaches

Table 5.12 includes two additional results as baselines that we computed using the following two solutions: (i) Twitter's metadata, which the system itself provides with each tweet, but it does not recognize Basque, Catalan, and Galician, and (ii) TextCat, a state-of-the-art n-gram-based language identification system developed for formal texts, which can deal with the six languages considered in the task. Note that TextCat was run after cleaning up the tweets by removing hashtags, URLs, and user mentions, as well as lower-casing the texts. The low performance of both

#	TEAM	P	R	F1
1	ELiRF @ UPV II	0.891	0.886	0.889
2	ELiRF @ UPV I	0.897	0.880	0.888
3	Citius-imaxin I	0.891	0.871	0.881
4	RAE II	0.884	0.869	0.877
5	RAE I	0.882	0.866	0.874
6	UB/UPC/URV	0.887	0.852	0.869
7	CERPAMID I	0.856	0.838	0.847
8	Citius-imaxin II	0.847	0.828	0.837
9	CERPAMID II	0.832	0.815	0.824
10	LYS @ UDC I	0.807	0.790	0.798
11	IIT-BHU	0.781	0.790	0.786
12	LYS @ UDC II	0.653	0.639	0.646

Table 5.10: Microaveraged performance results for all the submissions to the constrained track, sorted by F1 measure.

solutions, with F1 values below 0.5, emphasizes the difficulty of the task, as well as the need for proper alternatives for social media texts.

Evaluation with Respect to Unresolved Issues

In line with our motivation to study three key unresolved issues in language identification, we now delve into the analysis of results by looking into the performance of the systems when it comes to these three aspects separately: (i) performance results by tweet length, (ii) performance results for monolingual and multilingual tweets, and (iii) performance between similar languages by looking at the confusion matrix.

(i) Evaluation by Tweet Length. Figure 5.1 shows the performance of the systems by tweet length. These results clearly show the tendency of language identifiers to classify with substantially higher accuracy the tweets with more than 60 characters; the performance of the systems progressively drops especially for tweets with fewer than 60 characters. The performance is dramatically lower for tweets as short as 20 characters or fewer. While this corroborates the findings in previous works on language identification, it shows that language identifiers can also perform accurately

#	TEAM	P	R	F1
1	Citius-imaxin I	0.898	0.878	0.888
2	ELiRF @ UPV II	0.839	0.854	0.847
3	ELiRF @ UPV I	0.820	0.802	0.811
4	Citius-imaxin II	0.806	0.788	0.797
5	LYS @ UDC II	0.792	0.776	0.784
8	CERPAMID I	0.767	0.751	0.759
7	LYS @ UDC I	0.749	0.733	0.741
9	CERPAMID II	0.733	0.718	0.726
6	UB/UPC/URV	0.715	0.701	0.708

Table 5.11: Microaveraged performance results for all the submissions to the unconstrained track, sorted by F1 measure.

System	P	R	F1
Twitter	0.457	0.498	0.463
TextCat	0.586	0.480	0.447

Table 5.12: Performance results of baseline approaches using existing tools and resources, which enable comparison with the submitted systems.

for long tweets. Even though there is still room for improvement with long tweets, the main challenge remains in the correct identification of language for short tweets.

(ii) Evaluation for Monolingual and Multilingual Tweets. Figure 5.2 shows the results that the systems achieved for monolingual and multilingual tweets. As expected, the language identifiers performed substantially worse for multilingual tweets than for monolingual tweets. It is worth mentioning again that only two of the seven participants produced multilingual labels in their outputs, which means that for the other five systems, the evaluation is performed assuming that they will always miss at least one of the language in the multilingual ground truth. The two systems that produced multilingual labels, ELiRF and RAE, did obtain the best performance scores for the subset of multilingual tweets, with 0.453 and 0.39 F1 measure, respectively. Still, others who did not produce multilingual labels were not far from them, such as IIT-BHU achieving 0.37 F1 measure, and CERPAMID achieving 0.356. Even if the systems who considered multilingualism as a possible

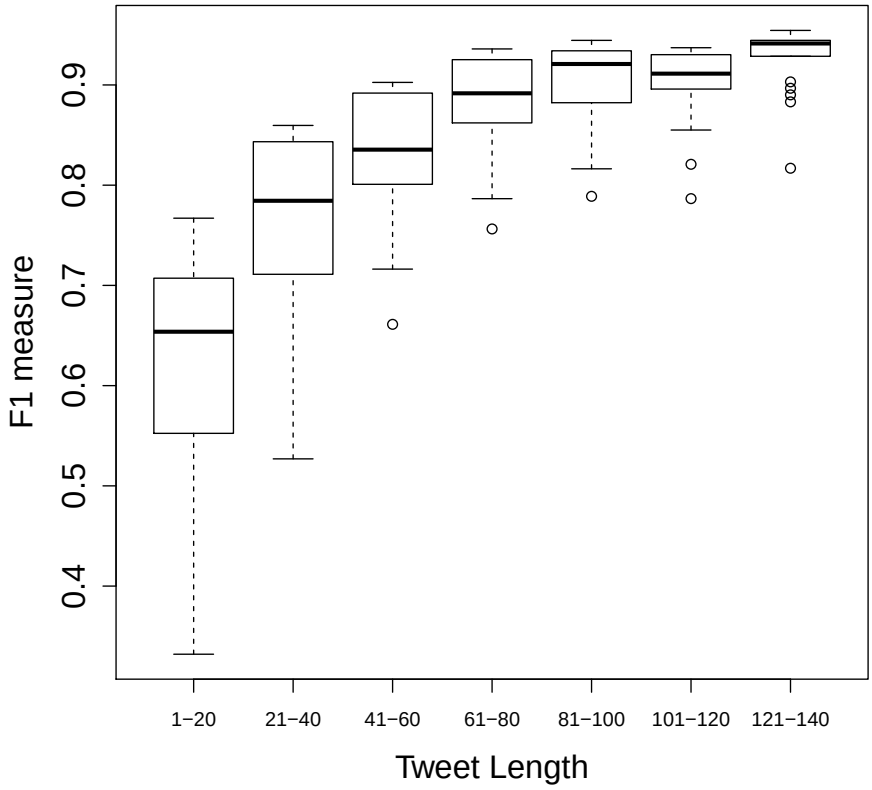


Figure 5.1: F1 scores achieved by submitted systems for different tweet lengths (tweet lengths measured as character counts after removing hashtags, user mentions, and URLs)

output performed better, the relatively small difference with respect to other systems shows the difficulty of dealing with these cases.

Despite the unsurprising fact that the systems performed worse for multilingual tweets, this analysis does, however, help us quantify the difference in terms of F1 measure between monolingual and multilingual tweets, where the classification of the former is about 20% more accurate than the latter. This posits an important drop in performance when tweets are of multilingual nature, which emphasizes the importance of properly dealing with multingual tweets, and leaves a challenge open for future research in tweet language identification.

(iii) Evaluation by Language, Focusing on Similar Languages. Figure

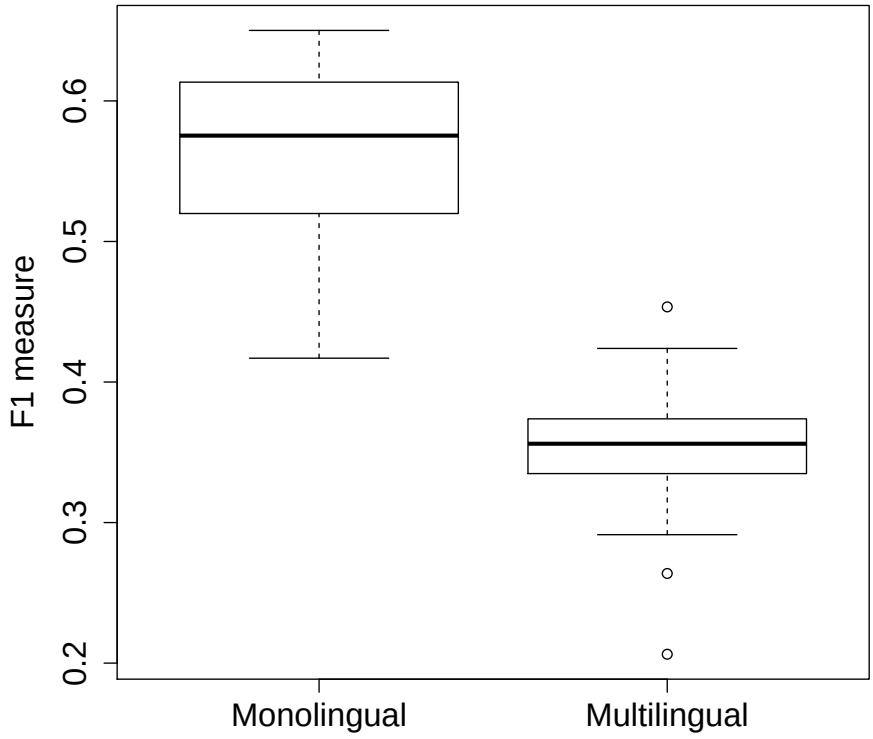


Figure 5.2: F1 scores achieved by the submitted systems for monolingual and multilingual tweets.

5.3 summarizes in a boxplot the distribution of precision values achieved by the 21 submitted systems for the different categories. These boxplots enable the visualization of quartiles in the ranked list of performance values; the bottom and top edges represent 0% and 100% percentiles, the bottom and top of the box represent the 25% and 75% percentiles, and the middle line represents the median, which allows to compare the distributions of performances for each language. It can be seen that the systems performed poorly especially for Galician (gl); this can be due to its similarity to Spanish (es) and Portuguese (pt), and its little presence in the corpus. Because of this similarity, and of course the cultural proximity where users tend to mix up spellings, the system might have had a tendency to picking

the most popular of the languages in these cases as output. The systems performed better for the rest of the languages, but still surprisingly there is a high variation of performances for Basque (eu), where we can see that some of the systems performed poorly. This is rather surprising given that Basque is very different from the rest of the languages, being an isolate language. A closer look at the errors by the lowest performing systems for Basque shows that these systems have a tendency towards picking the prevalent language (Spanish) for languages that have low representativity in the training set, such as Basque and Galician. Other systems, however, did better in dealing with the imbalance of the data, distinguishing what should be easier to distinguish from the rest of the languages, which is the case of Basque. Galician has, therefore, two challenges, its high similarity with respect to Spanish and Portuguese, as well as the small presence in the training set and the dataset. It also stands out that all the systems performed very well for Spanish, being this the majority language with over 60% of the tweets in the corpora.

Figure 5.4 complements the analysis with recall values achieved by the systems for the different languages. It can be seen that recall is especially low for undeterminable tweets as well as for Galician tweets. This highlights the difficulty of language identification systems to distinguish these cases from others; in the case of Galician, it is difficult to distinguish it from Portuguese and Spanish due to their much higher presence, and in the case of undeterminable tweets, it is a challenge to be able to determine that a tweet is not in any of the languages considered by the task, especially because the training set might not have or may have very few tweets in that specific language. Moreover, the recall is also slightly lower for Basque. Even if it is very different from the rest of the languages and hence reasonably easier to identify, its small presence in the training set harms the performance of some of the systems.

Figure 5.5 enables more detailed visualization of precision and recall values achieved by the systems for Basque and Galician, which as we mentioned above have proven challenging. These two charts show high diversity in the performance of the different systems, with few systems achieving a competitive balance of recall and precision values. The two systems performing best in these two cases, ELiRF for Basque and Citius-imaxin for Galician, have also achieved the best performances for the two tracks of the shared task.

Table 5.13 shows a confusion matrix comparing the ground truth and the aggregated outputs of all the systems for monolingual tweets, which allows us to analyze the extent to which the language identifiers tend to confuse between similar languages. To do this analysis, it is important to consider the bias derived from the

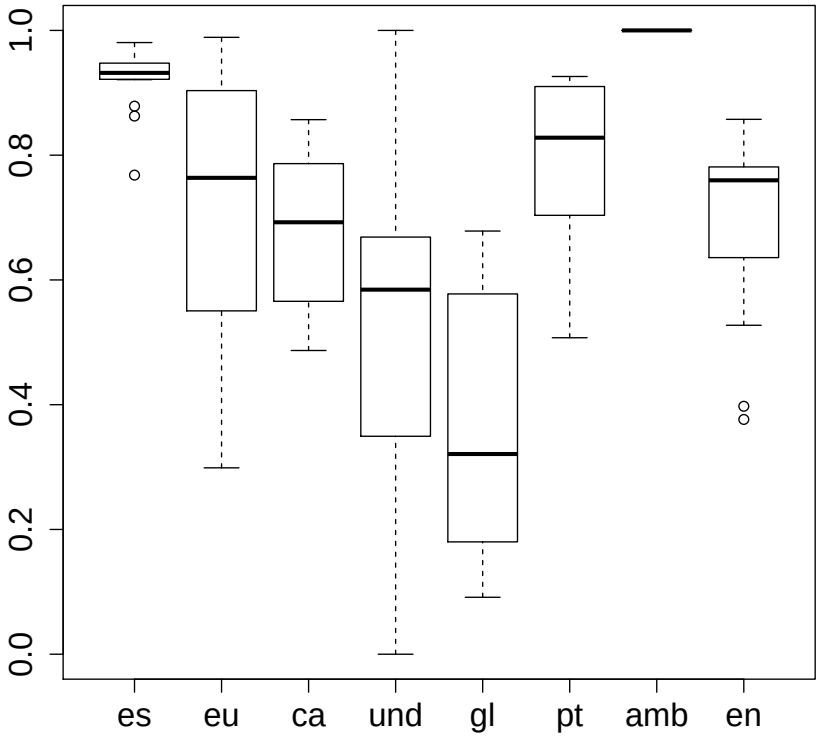


Figure 5.3: Distribution of precision scores by language for the 21 submitted systems, including results for both the constrained and the unconstrained tracks.

skewed distribution of tweets (a majority of them in Spanish) in both the training and test datasets. If we do not consider Spanish, Galician language tends to be mostly confused with Portuguese (12.7% errors from the total decisions), which is its closest linguistically related language. Similarly, besides Spanish, Portuguese is confused with Galician (3.2%) more often than with Catalan (1.3%), English (1.1%), or Basque (0.5%). In the case of Spanish, it is mostly confused with the other three Romance languages: Galician (3.5%), Catalan (2.4%), and Portuguese (2.2%), setting aside less related languages, namely English and Basque. Despite this was an anticipated and largely expected outcome, it emphasizes that language similarity is an important issue that reveals the shortcomings of state-of-the-art

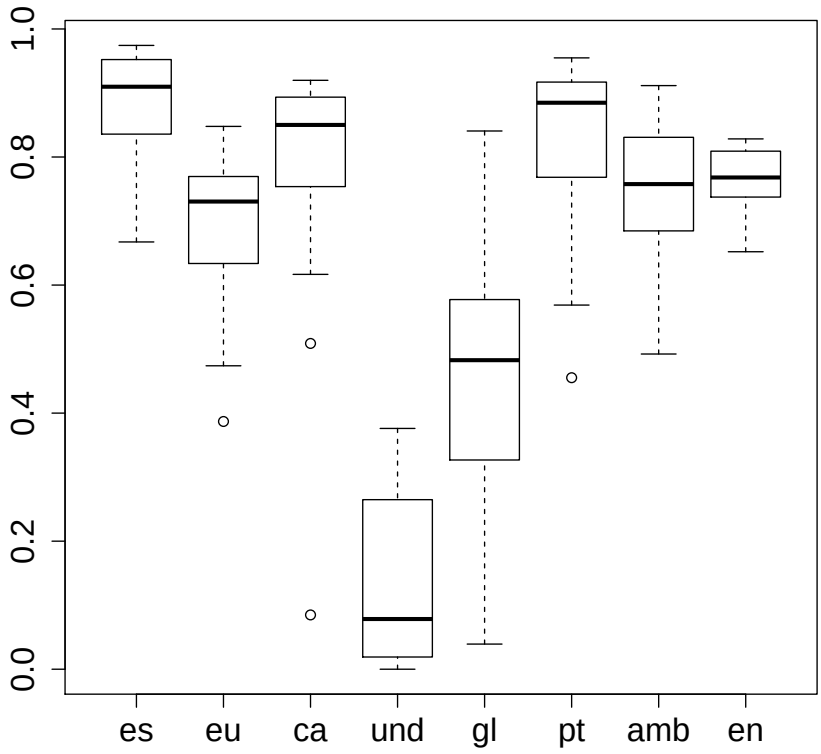


Figure 5.4: Distribution of recall scores by language for the 21 submitted systems, including results for both the constrained and the unconstrained tracks.

language identifiers.

Misclassified Tweets

Now we look at the errors produced by the participating systems, as well as the benefit that they could obtain from one another by combining them into a single classifier. First, we combined the output of all the participating systems by majority vote, so we we can obtain a single output for each tweet by aggregating the outputs. Table 5.14 shows the performance of a system that would combine all the systems, and compares it to that of the best system developed by ELiRF @ UPV. The

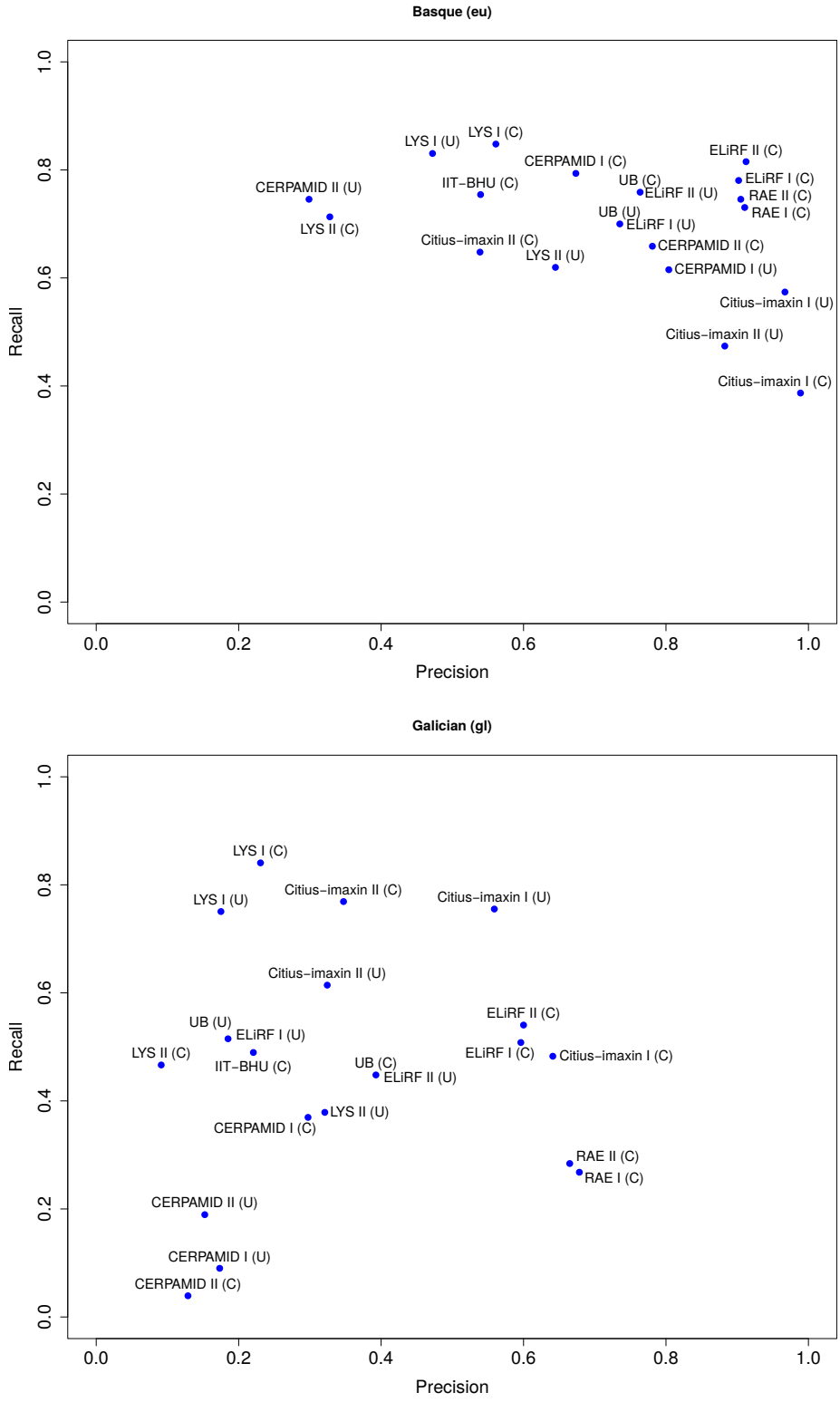


Figure 5.5: Scatter plots showing the precision and recall values for the 21 submitted systems, for tweets in Basque and Galician.

	pt	**ca**	**es**	**en**	**eu**	**gl**	**other**	**multi**	**und**
pt	**83.6**	1.3	8.6	1.1	0.5	3.2	0.6	0.6	0.5
ca	1.6	**81.8**	12.5	1.3	0.5	0.8	0.8	0.5	0.3
es	2.2	2.4	**88.1**	1.7	0.7	3.0	0.7	0.4	0.7
en	1.1	2.9	4.6	**87.4**	0.9	0.8	1.0	0.5	0.8
eu	0.9	1.2	10.1	2.9	**76.4**	1.5	1.9	1.3	3.8
gl	12.7	1.5	33.6	0.8	0.3	**47.2**	2.1	1.4	0.4

Table 5.13: Confusion matrix showing the percentage of cases whose ground truth is the language in the column and has instead been classified as the language in the row by the systems. The sum of the values in a row add up to 100%. The values in the diagonal (in bold) represent correct classifications, while the rest represent the percentage of deviations from the language in the column to the language in the row.

combined system can outperform the best system, with slightly better results when both macro-averaging or micro-averaging the performance values.

Among the 18,423 tweets considered for evaluation in the test set, we identified 600 tweets that were not guessed correctly by any of the submitted systems. Next, we look at some of these tweets, which allows us to analyze examples of the most challenging cases.

Multilingual tweets with low presence of one of the languages. This kind of tweets are probably the most difficult to deal with because both user intent and cultural habits are combined. Code-switching phenomena are a constant on social media as we have observed in this work. These tweets can often present a challenge even for human annotators. If we take a look at example 1, the use of the verb in Catalan (ets = you're) denotes the intent of the user to write in Catalan. The second part however, is written in Spanish (lo mejorcito = the best). In example 2 the tweet is mainly written in Basque, but the writer ends with a Spanish expression (si o si = come what may).

Example 1
@username ets lo mejorcito

[most systems categorized as "es", while it actually combines "ca+es"]

Example 2
Dutxita eta gerrate zibilakin gaur bukatubiou si o si

[most systems categorized as eu, but it actually combines "es+eu"]

Macro-averaged			
System	**P**	**R**	**F**
Meta-learning	**0.832**	**0.757**	**0.768**
Best system (ELiRF @ UPV II)	0.825	0.744	0.752
Micro-averaged			
Meta-learning	**0.910**	**0.892**	**0.901**
Best system (ELiRF @ UPV II)	0.891	0.886	0.889

Table 5.14: Results of a meta-learning approach that combines the output of all participating systems by majority vote, compared with the results for the best system in the shared task.

Lack of identification of NEs. In some cases, tweets written in a certain language contain NEs which are written in their original language. In the example below, even though the tweet is written in Galician, it contains the name of a TV show in Spanish ("Hay una cosa que te quiero decir"). Not identifying the NE leads the systems to confusion.

Example 3
Ese neno de "Hay una cosa que te quiero decir" é puuro amorr

[most systems categorized as "es", while it is actually "gl", but the NE (quoted) is indeed "es" and confuses the classifiers]

Difficulty to identify undeterminable ("und") tweets. In some cases, due to lack of clarity, or because of the brevity of some tweets, not even a human can determine what language is used in a tweet. For instance, in the examples 4 and 5, it is hard to determine what the meaning of the tweets is without additional context, which led to the manual annotation as undeterminable ("und"). It is a challenge, though, for a language identifier to realize what these cases are. The systems generated very diverse predictions for these cases, suggesting that there is no strong similarity with any of the languages.

Example 4
@user skiada top!

Example 5
Tu + eu = uiui

Difficulty to identify tweets written in other languages. The correct prediction on these cases should be "other", as the tweets are written in a language different from those considered in the task. Examples 6 and 7 show two tweets written in Dutch. Participant systems had no language model for that language, and therefore were unable to determine what language it was, and even to determine that it is not one of the languages under consideration. In general, our intuition is that this kind of tweets obtain similarity scores that suggest that they are not far enough from the other language models so as to be regarded as "other".

Example 6
#CaminoVascoDelInterior We zijn in Spanje : eerste mojón met schelp in Irún {URL}

[the systems generated very different predictions, while a tweet in Dutch should be marked as "other"]

Example 7
Naar bed naar bed zij duimelot

[the systems generated very different predictions, while a tweet in Dutch should be marked as "other"]

The list above summarizes the most frequent types of categorization errors when we look at the tweets misclassified by all of the systems analyzed in this work. Other common errors, such as deviations between similar languages, do not appear in this list given that they are usually guessed correctly by at least one of the systems.

5.7 Discussion

In this work, we describe and release a benchmark dataset and evaluation framework for tweet language identification. Through the shared task we organized to encourage researchers to submit the results of their language identification systems applied to this dataset, we looked at content-based tweet language classification approaches. The study of other features that a social network like Twitter can offer, such as user metadata, are not within the scope of this work and are left for future work.

5.7.1 Performance of Tweet Language Identification Systems

We were especially interested in this case in studying state-of-the-art approaches for language identification in a new scenario like Twitter. However, we do believe

that the use of features inherent to the social network can be of help for a language identifier, especially for adding context when the content is insufficient. We believe that the study of additional user-related features can help (i) when tweets are very short, looking for instance at previous tweets posted by a user, which might reveal what language the user uses most, and (ii) when two similar languages need to be distinguished, for instance looking at the location of a user, which might help identify the language(s) that are likely used in that location.

As we have shown, multilingualism is also a challenging issue in short texts like tweets. Further exploiting the social network, one could look at the historical tweets of a user to first list the languages that a user is likely to use, to then determine if the user has used a combination of those in new tweets; this involves having to look at more tweets from a user though, which is costly in terms of API accesses required, and might not always be feasible.

Regarding NEs, none of the participants tried to incorporate NER capabilities to their system, but it could have been useful as shown in some of the misclassified examples above. However, the use of NEs for this task is not trivial. For the shared task, our choice was to ignore NEs when annotating the language. While some NEs can be good hints about the language of the user, such as place names because they are usually translated into the corresponding language (e.g., Donostia (eu) vs. San Sebastian (es), other NEs however tend to be used both in their original form and in their translated form, e.g., Spanish tweeters use both "Game of Thrones" (en) and "Juego de Tronos" (es).

While the shared task we conducted, as well as the analysis of the submitted systems we discuss here, do not consider other social network features beyond a tweet's content, the dataset we created and released to the scientific community does allow to collect and incorporate these extra features for further analysis.

5.7.2 Comparing Errors between Human Annotators and by Language Identification Systems

Throughout the paper, we have studied both the performance of human annotators as well as that of the automatic language identification systems. The human annotations have been assessed by having two annotators annotate a 10% sample of the whole, while the automatic systems have been evaluated comparing against the manually defined annotation as the ground truth. Both evaluations have shown, to some extent, a similar tendency; both humans and systems struggled to identify the language of short tweets as well as the languages in multilingual tweets, and also found it difficult

to distinguish similar languages. Still, there are a number of differences between the performances of humans and systems, which helps us set forth a set of objectives for future work.

Length of tweets: while human annotators performed lower for very short tweets of less than 20 characters, the performance was quite consistent for other lenghts. The systems, though, showed a progressive decay in performance as the tweets are shorter, experiencing a significant drop in performance for tweets of less than 60 characters. While improving the performance of the language identification for tweets of less than 20 characters might not be viable, we believe that there is still room for improvement for tweets between 20 and 60 characters, which humans could label as accurately as longer tweets.

Multilingualism: multilingual tweets have proven challenging both for human annotators and for language identification systems, with a significantly lower performance than for monolingual tweets. However, only two of the participants in the shared task developed systems that would ever output a multilingual label, which makes our analysis in this aspect still inconclusive enough so as to conclude the extend to which it can be improved. The better performance of the two systems that implemented multilingual outputs over the rest of the systems, however, does encourage to perform further research. We believe that testing more multilingual systems would help extend the analysis of classifying multilingual tweets.

Similar languages: the confusion between similar languages occurred differently for human annotators, given that each annotator had to deal only with tweets from a specific region, which means that there could be rarely confusions between Spanish and Portuguese, because they usually appear in different regions. Still, for one of the most common errors in our dataset, i.e., confusions between Galician and Spanish, human annotators performed much better, and language identification systems missed as many as 33.6%. In the latter case, the performance worsens owing to the fact that Galician has fewer instances in the training set, which also occurred with Basque, a language which is very different from the rest, but its low presence occasionally harms the performance of classifiers. Better dealing with similar languages, as well as better managing languages with fewer instances in the training set, are certainly two of the key aspects to look at in the future.

5.7.3 Contributions and Limitations of the Shared Task

Through the organization of TweetLID as a shared task, we have fulfilled most of the objectives we set forth at the beginning of planning this work, and we expect that our

contributions will help pave the way to researchers aiming to study tweet language identification in the future. However, we have also identified a set of limitations in the shared task.

On the positive side, we believe that TweetLID has managed to attract a good number of participants, who have submitted a diverse set of systems. This has enabled a quite complete analysis of language identification systems applied to tweets, as well as the identification of main directions for future research. This has been possible thanks to the creation of an annotated corpus of tweets that meet the main characteristics we seeked, as well as the definition of the evaluation methodology. This corpus will in turn enable further research in the future. Thankfully, Twitter's newly revised terms of service allows us to release the content and all the metadata of the tweets, which will guarantee that whoever is interested will be able to retrieve the complete dataset, which will not shrink over time.

On the other hand, one of the weak points of the systems submitted to the shared task has been the limited attempt at dealing with multilingual tweets. In fact, only two of the seven participants produced language identification systems that would ever return a multilingual label as output. While it has not been possible to test additional multilingual systems in this shared task, it would have been useful to have more such systems participating, and would be ideal to have in a future shared task. Moreover, even if it was originally restricted in the definition of the shared task, we have not let participants to make use of tweet metadata to identify languages, which would also be wise to study in an upcoming shared task.

Last but not least, it also makes it extra challenging to organize the shared task the fact that Twitter's terms of service did not allow us to share tweet content with the participants. Instead, we gave them the list of tweet IDs, which they used to retrieve the content of the tweets themselves by accessing Twitter's API, which leads to each participants having slightly different sets of tweets due to some tweets becoming unavailable over time. The updated terms of service would enable, however, to share the content with the participants of future tasks.

5.8 Conclusion

The Twitter dataset with nearly 35,000 tweets with language label manually annotated has enabled us to study currently unresolved issues in language identification. These include the following three issues: (i) short texts provide very little context to determine the language of their content, (ii) multilingual texts

make it more difficult identify the presence of the different languages, and (iii) similar languages are very difficult to distinguish from each other. Our Twitter dataset provides a suitable resource to study the aforementioned issues, which we have put into practice and analyzed through the TweetLID shared task where seven participants submitted the output of their language identifiers. Thanks to the development of this dataset and the shared task to assess the performance of different systems, we have come up with an evaluation methodology that can be of help to researchers in the field.

Our dataset included the five top languages of the Iberian Peninsula –Spanish, Portuguese, Catalan, Basque, and Galician– as well as English. This has allowed participants to compare their systems with four romance languages that share similarities with one another, and two more languages that are substantially different from the rest, i.e., English and Basque. The participants have applied state-of-the-art language identification techniques designed for other kinds of texts such as news articles, as well as adapted approaches that take into account the nature of the brevity and chatspeak found in tweets. Still, the performance of the systems posits the need of further research to come up with more accurate language identification systems for social media. Some of the key shortcomings that the shared task has brought to light include the need for a better choice of external resources to train the systems, the low accuracy of the systems when dealing with underrepresented languages which are very similar to others –as occurred with Galician here–, and the inability to identify multilingual tweets. Future work on tweet language identification should look into these issues to develop more accurate systems.

CHAPTER 6

Microtext Normalization Benchmark

TweetNorm: A Benchmark for Lexical Normalization of Spanish Tweets

Iñaki Alegria[1], Nora Aranberri[1], Pere R. Comas[2], Víctor Fresno[3], Pablo Gamallo[4], Lluis Padró[2], Iñaki San Vicente, Jordi Turmo[2], Arkaitz Zubiaga[6]

[1] IXA. UPV/EHU, [2] UPC, [3] UNED, [4] USC, [5] Elhuyar, [6] University of Warwick

The language used in social media is often characterized by the abundance of informal and non-standard writing. The normalization of this non-standard language can be crucial to facilitate the subsequent textual processing and to consequently help boost the performance of natural language processing tools applied to social media text. In this paper we present a benchmark for lexical normalization of social media posts, specifically for tweets in Spanish language. We describe the tweet normalization challenge we organized recently, analyze the performance achieved by the different systems submitted to the challenge, and delve into the characteristics of systems to identify the features that were useful. The organization of this challenge has led to the production of a benchmark for lexical normalization of social media, including an evaluation framework, as well as an annotated corpus of Spanish tweets –TweetNorm_es–, which we make publicly available. The creation of this benchmark and the evaluation has brought to light the types of words that submitted systems did

best with, and posits the main shortcomings to be addressed in future work.

Published in *Language Resources and Evaluation*, 49(4):883–905, Dec 2015. ISSN 1574-0218. doi: 10.1007/s10579-015-9315-6

6.1 Introduction

With its evergrowing usage as a microblogging service, Twitter has become a ubiquitous platform where users continuously share information in a real-time fashion. Information is posted by users in the form of *tweets*, which are characterized by their brevity, restricted by Twitter's 140 character limit, and which often lack correct grammar and/or spelling. This posits the need for a process of lexical normalization of these tweets as a key initial step for subsequently applying natural language processing (NLP) tools such as information extraction, machine translation and sentiment analysis. Even though research on lexical normalization of tweets is still in its infancy, early studies have shown that it can indeed boost performance of NLP tools that work on tweets (Wei et al., 2011). While lexical normalization of SMS and tweets in English has attracted the interest of a community of researchers (Han and Baldwin, 2011; Han et al., 2013a; Liu et al., 2012), little has been studied for this kind of short texts written in other languages such as Spanish.

A lexical normalization system takes a natural language sentence as input, and consists of the following two stages: (i) non-standard word detection, which identifies the words from the input sentence that need to be normalized, and (ii) candidate selection, which selects the alternative word as the normalized output. As a result, the objective of a lexical normalization system is to output a modified version of the input sentence, such that non-standard words have been normalized. Both stages are crucial to build an accurate normalization system, since a wrong decision in the first step as to whether a word needs to be normalized, will lead to a bad candidate selection in the subsequent step and thus to an inaccurate normalization of the word. This inaccurate normalization can be twofold. On one hand, mislabeling a word as "non-standard" will lead to the wrong detection of a candidate and, on the other hand, mistakenly identifying a non-standard word as being correct will skip the subsequent candidate selection stage when it is really needed.

In order to motivate additional research in the field, we organized the Tweet-Norm 2013 shared task[1] held at the 29th Conference of the Spanish Association for Natural Language Processing[2] (SEPLN). The goal of the shared task was to create a benchmark for lexical normalization of microtexts written in Spanish, including both a robust evaluation framework, and an annotated corpus to perform the experiments with. In this shared task, participants were provided with such corpus and evaluation guidelines, and were asked to normalize a set of tweets containing several non-standard word forms each.[3]

We created a corpus of tweets annotated with normalized variations, which we released to participants of the shared task for benchmark evaluation purposes. The creation and distribution of a benchmark corpus provided a common testing ground, which enabled us to compare performance of participants and to identify the main advantages and shortcomings of each participating system. This paper makes the aforementioned corpus publicly available with the aim of attracting researchers and practitioners to develop new normalization approaches while making use of a common evaluation setting. We describe the methodology followed for the collection of tweets and the generation of the annotated corpus that has been put together in the resulting *TweetNorm_es* corpus. We also present a detailed analysis of the results of the shared task, delving into the performance of each system and breaking down performance values into word categories, including common words, onomatopoeias, entities, and others. This detailed analysis allows us to shed some light on the types of approaches that can be of help to build accurate normalization systems, and to set forth the main shortcomings that need to be addressed in future research in the field.

The corpus can be used, modified and redistributed under the terms of the CC-BY license.[4]

6.2 Related Work

Twitter is being used increasingly as an information source for NLP research in multiple tasks. These include sentiment classification (Go et al., 2009; Jiang et al., 2011), topic modeling (Lin et al., 2011; Hong and Davison, 2010), and summarization

[1]Details about the workshop can be found at `http://komunitatea.elhuyar.org/tweet-norm/`

[2]`http://nil.fdi.ucm.es/sepln2013/`

[3]The term "ill-formed" has also been used in the literature to refer to these non-standard word forms. We opted for the term "non-standard word form" because some of the words that fall into this category, such as abbreviations or acronyms, are not necessarily misspellings.

[4]`http://creativecommons.org/licenses/by/3.0/legalcode`

(Inouye and Kalita, 2011; Chakrabarti and Punera, 2011), among many others. However, Twitter poses an unprecedented problem for NLP research; the fact that users tend to shorten their texts to make it fit into a tweet, often with the extra challenge of posting from a mobile device, and the occasional misspellings and typos they introduce, leads to the creation of short texts characterized by a non-standard language, which makes the analysis of texts more challenging. Eisenstein (2013) outlines the challenges posed by the *bad language* that characterizes the Internet, and surveys two of the most popular directions in which the NLP community has tackled this issue: normalization and domain adaptation.

However, the lexical normalization of tweets is still in its infancy as a research field. Some of the early work in the field by Han et al. (2011; 2013a), comparing different approaches for lexical normalization, found that approaches based on language models perform significantly worse than dictionary lookup methods. This is likely due to the fact that the lexical context in Twitter data is noisy; on many occasions, Out-of-vocabulary (OOV) words can co-occur with user mentions, hashtags, URLs, and even other OOV words, which can produce poor context information. They also observed that well-known methods for normalization in other domains suffer from the poor performance of lexical variant detection, which worsens the effectiveness of existing techniques to the context of Twitter and social media. The best system by Han and Baldwin (2011), which is mainly based on dictionary lookup, achieved an F-score of 75.3% in a partial evaluation that focused only on the candidate selection step, assuming that the previous step for non-standard word detection was perfect.

Motivated by the performance of the dictionary-based normalization system, in a later study Han et al. (2013a) enhanced their system by using information from both word distribution and string similarity to build normalization lexica with broader coverage. They reported an F-score of 72.3% when dealing with the whole task, i.e., including both the OOV detection and normalization steps. These results can be considered to be the state-of-the-art for normalization of English tweets. Their case study, albeit focused on English tweets, is straightforwardly applicable to other languages, given that they defined a generalizable research methodology for this kind of tasks. In the present work, we relied on their methodology to define the corpus annotation guidelines, as well as to set up the shared task. However, there are a few differences that we introduced in our case:

- As we said above, most previous work assumed perfect OOV detection, and focused on the subsequent candidate selection step. Instead, the shared task held at Tweet-Norm 2013 considers both the detection of OOVs, and the

normalization as a single process. To our knowledge, only Han et al. (2013a) have proposed such an integral solution.

- Different from both (Han and Baldwin, 2011; Han et al., 2013a), the shared task at Tweet-Norm 2013 also considers multiwords in the lexical normalization process. We considered one-to-many correspondences (e.g., *imo* → *in_my_opinion*), and so the submitted systems had to deal with multiwords.

Using several corpora, including the one described above, Liu et al. (2012) propose a normalization system, and especially focus on exploring the coverage achieved by this system when applied to SMS and Twitter data. They propose a cognitively-driven normalization system that integrates different human perspectives into the normalization of non-standard tokens, including enhanced letter transformation, visual priming, and string/phonetic similarity. Results show that the presented system achieves over 90% word-coverage across all datasets.

Others have also performed a preliminary tweet normalization step prior to the main task. For instance, Wei et al. (2011) perform a 4-step normalization of English tweets before running their topic detection system: (i) OOV word detection, (ii) slang word translation, (iii) candidate set generation, and (iv) candidate selection. They performed an *in vivo* evaluation of their system, looking instead at the performance boost of the system presented to the TREC 2011 microblog search track. They found the normalization system to be effective, providing a slight improvement to the results, although a more comprehensive normalization system could do even better. Similarly, Liu et al. (2013) used a tweet normalization system as the initial step of their system for named entity recognition. They use statistical learning algorithms, trained with the pairs provided by Han and Baldwin (2011), as well as 1,500 more pairs which were compiled manually. They obtained an F-score of 60.5%.

Others have opted for making use of large-scale data collections to train their normalization systems. Examples include (Beaufort et al., 2010), who tackle the task of normalizing SMS texts using the noisy channel model, very common in speech processing, or (Kaufmann and Kalita, 2010), who feed a statistical machine translation model with tweets, to turn them into standard English. Another example is the work by Ling et al. (2013), who make use of self-translation from Twitter and Sina Weibo in order to obtain large-scale (albeit noisy) normalization examples. The Mandarin version is automatically translated back into English, and then two versions are available in English: the original (not normalized) and the noisy translation from the equivalent tweet in Mandarin (noisy normalized). Then, they use the SMT

framework for learning the normalization patterns. However, despite research in this direction, in this work we are interested in avoiding the need for large-scale training data, as this tends to be costly. Moreover, no such resources are available to the best of our knowledge for microblogs, in particular for languages other than English. Hence, we wanted to define the tweet normalization task assuming the limited availability of training data, and thus allowing participants of the task to focus on the algorithms and external resources that can be of help.

Wang et al. (2013) tackled a related task for microtext normalization task, in this case in Chinese. Instead of normalizing the spelling of words, they studied the translation of Chinese texts into their formal alternatives. This is especially important given that machine translation systems tend to mistranslate informal words when translating for instance into English. The authors studied first the linguistic phenomenon of informal words in the domain of Chinese microtext, and presented then a method for normalizing Chinese informal words to their formal equivalents. The task is formalized as a classification problem and proposes rule-based and statistical features to model three channels or phenomena (i.e., phonetic substitution, abbreviation and paraphrase) that identify connections between formal and informal pairs. They created a corpus for evaluation purposes, which was annotated through crowdsourcing, and reported a precision score of 89.5%.

To the best of our knowledge, previous efforts have focused on language specific approaches for English tweets, there is limited work for Chinese in a related task, and no work has dealt with tweets written in Spanish. Our work intends to fill this gap by tackling the normalization for Spanish tweets, defining a benchmark for evaluation. To the best of our knowledge, this is the first work that deals with lexical normalization of non-English tweets. As a related effort for tweets in Spanish, Villena Román et al. (2013) organized a shared task focused on sentiment analysis. Costa-Jussà and Banchs (2013) and Oliva et al. (2013) have also worked on normalization of SMS texts in Spanish.

6.3 Corpus

In this section, we describe the process we followed to collect and sample the tweets, which were manually annotated. First, we describe the API settings defined to collect the tweets, and then explain the preprocessing step carried out to prepare the data for manual annotation.

6.3.1 Tweet Dataset

Among the Twitter APIs[5] for tracking and collecting tweets, we opted for geolocated tweets, whose metadata include the coordinates of the location each tweet was sent from. Twitter's API allows the user to select tweets sent from a pre-determined geographic area. Making use of this feature, we chose an area within the Iberian Peninsula, taking out regions where languages other than Spanish are also spoken. We found this approach to be highly effective when it comes to gathering large numbers of tweets in Spanish. Thus, the selected geographic area forms a rectangle with Guadalajara (coordinates: 41, -2) as the northeasternmost point and Cádiz (coordinates: 36.5, -6) as the southwesternmost point. The collection of tweets gathered on April 1-2, 2013 amounts to 227,855 tweets. From this large dataset, we created two random subsets of 600 tweets each, which were shared with participants, one as a training set, and the other as a test set for final evaluation purposes. The rest of the dataset was also shared with participants, with no manual annotations, which they could use for setting up their unsupervised normalization systems.

6.3.2 Preprocessing

We used the FreeLing[6] language analysis tools (Padró and Stanilovsky, 2012) for the identification of OOV words from tweets. We used some of the basic processing modules included in this library to tokenize and analyze tweets. A token was ultimately considered to be an OOV when none of the modules identified it as an in-vocabulary (IV) word.

We used the following modules to process the tweets:

- tokenizer.
- usermap.
- punctuation detection.
- number detection.
- date detection.
- morphological dictionary (with affixes handling).

[5] `https://dev.twitter.com/docs/api`
[6] `http://nlp.cs.upc.edu/freeling`

- quantitites detection.

These modules were set up with their default configuration, except in the following cases, for which we detail the changes we made:

- `tokenizer`: The rules of the tokenizer were tuned to keep usernames (`@user`), hashtags (`#hashtag`), e-mail addresses, URLs, and the most common emoticons as single tokens.
- `usermap`: We also enabled the `usermap` module (disabled in the default FreeLing configuration), which checks if each token matches one of a set of regular expressions that are discarded from being considered as OOVs. These regular expressions help identify usernames, hashtags, e-mail addresses, URLs, and common emoticons.

Specific configuration files used for `tokenizer` and `usermap` were later included in the FreeLing distribution and are available at the project's SVN repository.

On the other hand, the following modules for morphological analysis were disabled:

- multiword detector (to avoid agglutination of several tokens into a single one).
- named entity detector (since we want to keep them as OOVs).
- lexical probabilities module (which includes a guesser that would assign at least one analysis to each word).

6.4 Annotation Methodology

During the annotation process, experts were asked to annotate the OOV words. They tagged each OOV word either as *correct*, *variant* or *NoES* (not in Spanish). For those cases deemed variant, they also provided the normalized spelling of the word along with the annotation. Standard word forms are derived from the RAE dictionary[7].

Three experts independently annotated each OOV word for the development set, and two of them participated in the annotation of the test corpus. We put together

[7]RAE, or *Real Academia Española*, is the institution responsible for regulating the Spanish language.

the annotations from the different experts by majority voting when possible, and by further discussing the correct annotation among the experts in case of ties. To facilitate the annotation process and subsequent discussions, we defined the following guidelines for each OOV word:

- When the word is included in RAE's dictionary: mark it as correct.
- When a well-formed word refers to a Named Entity (e.g., Zaragoza) or a loanword (e.g., Twitter): mark it as correct.
- When a word incorporates an emphatic or dialectal variation, it is misspelled, or lacks or misuses the acute accent: mark it as variation and provide the standard spelling (e.g., muuuuuuucho/mucho, kasa/casa, cafe/café).
- When more than one word is written together with no separation: mark it as variation and provide the standard spelling (e.g., asik/así_que, find/fin_de_semana).
- When a single word is split into smaller strings: mark all of them as variations and provide their standard spellings (e.g., im_presionante/impresionante, per_do_na_meeee/perdóname).
- When a word is unintelligible, a foreign word, or others (e.g., XD): NoES.

Note that the guidelines distinguish between "loanwords" and "foreign words". We consider "loanwords" those that, despite belonging to a language different from that of the tweet, have been assimilated by it and are used in everyday language (e.g. "tablet", "sandwich"). In contrast, "foreign words" are those that have not been assimilated, and therefore, do sound foreign (e.g. in the tweet "Igor gracias no sabia que te importara tanto joo tio!! pero es que eres mu feote sorry", "sorry" is a foreign word). Named entities, in turn, are treated separately.

These guidelines include the most common cases, but some of the cases we found were still not covered. In those cases, we met to further discuss each case in search of the most suitable solution.

Examples of uncommon cases not considered by the guidelines above include:

- `que estafa de tablet`
  [what a scam is this tablet]
  (in this case *tablet* is a loanword that is not included in the RAE dictionary yet, but the Spanish alternative *tableta* will incorporate this new meaning in the next release of the dictionary).

- `Me dispongo a ver Game of Thrones.`
  [I'm going to watch Game of Thrones]
  In this case the original name of the series was used instead of the Spanish translation *Juego de Tronos*).

One of the most challenging cases we identified during the annotation process was the normalization of abbreviations. In some cases, the context surrounding the abbreviated word in question is not sufficient to disambiguate its meaning and to identify the intention of the user. For instance:

- `cariiii k no te seguia en twitter!!!mu fuerte!!!..yasoy tu fan....muak....se te exa d menos en el bk....sobreto en los cierres jajajajas`
  [my dear i wasn't following you on twitter!!no way!!i'm so fan of you from now on....kisses... we miss you in the **bk**.... especially when closing hahaha]

where it is difficult to know what *bk* refers to with certainty. This addressee had seemingly a colleague at a place called *bk*, but there is little evidence to grasp its exact meaning without further research. The annotators ultimately chose *Burger King* as the variant, as the most likely choice for the acronym. In a few cases, OOVs could not be disambiguated and the annotators provided two alternatives. This includes cases where the gender could not be disambiguated from the abbreviated form –e.g., a tweet from the corpus contained the word *her*, which may refer to either *hermano* (brother) or *hermana* (sister).

The meaning of some onomatopoeias was also hard to grasp in some cases, which needed further discussion to come to an agreement among annotators. For instance:

- `me da igual JUUUM!!`
  [i don't care huuum!!]

6.5 Development and test corpora

Two collections have been generated from the initial corpus described in Section 6.3.1: the development corpus and the test corpus, which consist of 600 tweets each. A total of 775 and 724 OOV words were manually annotated respectively in both corpora.

Corpus	#OOV	0	1	2
Development	775	107	600	68
Test	662	98	531	33

Table 6.1: Distribution of the three OOV word categories (0, correct; 1, variant; 2, NoES) in the development corpus and in the final test corpus.

As required by Twitter API's terms of use,[8] we do not release the content of the tweets, but provide instead the user names and tweet IDs that enable to download the content of the tweets by using *Twitid*. *Twitid*[9] is a script that retrieves the content of tweets from the list of user names and tweet IDs.

Since we distributed the lists of tweets to participants following the method above, chances are that some tweets might have become unavailable. Some tweets may become unavailable as users remove their accounts, make them private, or delete the tweet. This may lead to participants having slightly different collections of tweets, which would affect the evaluation process. We figured this out by identifying the subset of tweets that were still available after all participants submitted their results. We found that 562 of the 600 tweets in the original test set were still accessible at the time. Thus, the initial set of 724 OOV words found in the initial test corpus was reduced to 662 due to the unavailable tweets. We relied on this slightly reduced set of tweets for the final evaluation.

Both datasets are publicly available[10] under the terms of the CC-BY license. The datasets include tweet IDs, user names and annotations. Note that participants had no access to the ground truth annotations of the test set during the test period.

Table 6.1 shows the distribution of the three OOV word categories (0, correct; 1, variant; 2, NoES) in both the development corpus and the test corpus. Note that the distribution of the three categories is similar in both corpora. This fact allowed the participants to develop their systems with a corpus that is similar to the test corpus.

[8] `https://dev.twitter.com/terms/api-terms`

[9] `http://komunitatea.elhuyar.org/tweet-norm/iles/2013/06/download_tweets.py`

[10] `http://komunitatea.elhuyar.org/tweet-norm/files/2013/11/tweet-norm_es.zip`

6.6 Tweet-Norm shared task

In this section, we first set out to describe the objective of the shared task. Then, we describe the characteristics of the systems that participated in the shared task, and finally present and analyze the results.

6.6.1 Objective and Evaluation Criteria

The Tweet-Norm shared task aimed at normalizing words unknown to the analyzer at the preprocessing step, such as abbreviations, misspellings, words with repeated characters, etc. Following the line of work of Han and Baldwin (2011) we focus on lexical normalization, whereas other phenomena such as syntactical or stylistic variants are left out of this task.

The goal of the task is to measure how accurate a system is at normalizing OOV words found in tweets. This goal does not involve the classification of the OOV words into different categories (0, 1 and 2, as described in previous section). Instead, the task focuses on identifying whether an OOV word needs to be corrected, and on providing the correct alternative when necessary. Participants had to determine if an OOV word should be deemed correct (e.g., new named entities, words in other language, etc.) or it should be assigned a normalized variation. We measured the accuracy of each system when performing the final evaluation as follows:

- **Correct**: if the OOV word is correct (category 0) or NoES (category 2) and the system does not provide any correction, or if the OOV word is a variant (category 1) and the word suggested by the system to normalize the OOV word is correct.
- **Incorrect**: otherwise.

In order to measure the performance of the systems, we relied on the precision score, defined as the number of correct responses of a system over the whole set of OOV words in the test corpus:

$$P(system_i) = \frac{\#\text{correct suggestions}}{\#\text{OOV words}}. \tag{6.1}$$

6.6.2 Short Description of the Systems

In this section we describe approaches utilized and the characteristics of the systems developed by the participants of the shared task. Next, we list the names of the participants, and describe the systems submitted by them. If it is not stated otherwise, descriptions correspond to the best submitted runs.

RAE (Porta and Sancho, 2013): RAE's system devises several rewriting rules that model specific spelling phenomena with very high precision. Other rules include edit distance and typing errors. These rules are compiled into finite state transducers that can be recombined in order to produce a confusion set for OOV words. OOV normalization candidates must occur in a lexicon composed of the DRAE dictionary for Spanish words, the BNC corpus for English words, and a list of NEs compiled from many sources. A language model (LM) decodes the word graph obtained, and determines the most probable sequence for each tweet. The system uses a 3-gram LM obtained from crawling 20k Spanish web pages.

Citius-Imaxin (Gamallo et al., 2014): This system produces two sets of candidates, using several lexical resources (including the DRAE dictionary, a normalization dictionary and names collected from Wikipedia). It also makes use of a set of predefined rules for three kinds of non-standard forms that need to be normalized, i.e., capital letters, repeated characters, and common misspellings. The system is trained with a LM developed from a news corpus gathered from RSS feeds, which is then used to select among the candidates. More precisely, the local context of each candidate is compared, by computing chi-square measure, against the LM which consists of bigrams of tokens found within a window of size 4 (2 tokens to the left and 2 to the right of a given token).

UPC (Ageno et al., 2013): To produce spelling alternatives for the OOV words, this system searches for similar variants in several gazetteers and lexica (Spanish, English, NEs, morphological derivates of Spanish words) using edit distance measure with several cost matrix: one for keyboard typos, one for phonetic similarity and normal edit distance. It also relies on hand-crafted regular expressions to detect onomatopoeias, acronyms and common shorthands. The final candidate is chosen with a weighed voting scheme: each producer (pair of lexicon and search method) is assigned a weight which is equivalent to its precision on the development data.

Elhuyar (Saralegi and San-Vicente, 2013): This system first generates all the possible candidates for the OOV words in a tweet, and then selects the combination of candidates that best fits a LM. For the generation of candidates, it combines common abbreviations, colloquial expressions, repeated characters, onomatopoeias

and typographical/orthographical errors. Reference lexica of normalized forms were generated from various resources. The LM used for the selection of correct candidates is built using SRILM based on bigrams obtained from Wikipedia articles and a news corpus from EFE[11], a Spanish news agency.

IXA-EHU (Alegria et al., 2013): This system uses hand-written rules (using *foma*) for the most common phenomena. These rules are incrementally applied. In a complementary way basic orthographic changes are learned and weighed using a noisy-channel model (*Phonetisaurus* tool). Frequencies on the full corpora of tweets after selection of correct words are used as LM. One-to-several correspondences are generated and filtered using a search engine.

Vicomtech (Ruiz et al., 2013): This system performs a first-pass correction using regular expressions and custom lists, which detects and corrects common errors and abbreviations. Then, it uses the edit distance as a measure, up to distance 2, to obtain spelling alternatives. The alternatives are ranked according to a LM and the edit distance scores. A postprocessing step detects NEs and corrects capitalization. The system is fed with a standard dictionary and NE lists obtained from multiple sources. The LM consists of 5-grams of film and documentary captions, although other alternative LMs were also tested.

UniArizona (Hulden and Francom, 2013): This system uses contextual phonological replacement rules in the form of transducers to convert the OOV into legitimate lexicon words. Two implementations of the same strategy are given: the first uses hand-crafted transformation rules (about 20), while the second automatically learns this rules using the noisy channel model (combining a transformation model and word frequency). The correction rules range from very specific to highly generic.

UPF-Havas (Muñoz-García et al., 2013): After separating Twitter metalanguage elements using regular expressions, the OOV words go through a pipeline with feedback loops. This consists of several stages: Spanish dictionary look-up using some spelling variants (case and accents), SMS dictionary look-up, repeated character correction, correction through an open-source spell-checker. The dictionary includes Spanish common names and NEs obtained from Wikipedia articles.

DLSIAlicante (Mosquera-López and Moreda, 2013): First of all, the system attempts to find the OOV word in the dictionary (aspell dictionary enriched with NEs) after applying heuristic rules to reduce character repetitions, the conversion of numerals into text and a table of abbreviations. If not successful, then the dictionary

[11] `http://www.efe.com/`

is indexed using the Metaphone algorithm and the closest spelling alternative is found using the longest common subsequence method. In the case of a tie, a 3-gram LM is used to select the final candidate.

UniMelbourne (Han et al., 2013b): In this approach, a collection of 280 million Spanish tweets is used as a source of IV words. For each OOV, the system generates spelling alternatives considering all words in the collection that have small edit distance either orthographically or under the Metaphone phonetic representation. The best correction is selected with a measure of distributional semantics similarity (KL divergence, (Kullback and Leibler, 1951)) in a window of two words. Additionally, a lexicon of slang and abbreviations has been hand-crafted.

UniSevilla (Cotelo-Moya et al., 2013): The normalization process begins with filtering the OOV words using a set of lexica covering the Spanish language, as well as small dictionaries built with Twitter vocabulary, emoticons and colloquial inflections. The OOVs are processed with hand-crafted rules (e.g., for character repetition or SMS language), spelling corrections based on edit distance, and a language identifier module. All this information is taken by a candidate selector based on confidence values that produces the final output.

UJaen-Sinai (Montejo-Ráez et al., 2013): This system uses a small lexicon of abbreviations and regular expressions that capture onomatopoeias. Then a spell-checker is used to produce spelling alternatives. The spell-checker lexicon is enriched with NEs (from Wikipedia and geographic information sources), popular Twitter jargon, neologisms and interjections in Spanish. The normalized candidate is selected with a unigram LM.

UniCoruña (Vilares et al., 2013): This system is based on a pipeline that applies rules that detect and normalize onomatopoeias, reduction of character repetitions, diacritic variations and a general purpose spell-checker. The system is trained with a SMS lexicon to enrich the spell-checker dictionary.

6.6.3 Results

Table 6.2 shows the accuracy results obtained by the 13 participants[12] in the shared task. The table includes an extra column with a second precision value for participants who submitted two runs. Besides the results of the participants, we also show two more results as references. Firstly, the *Baseline* would be the result of deeming all OOV words correct, therefore without suggesting any changes at all

[12]Out of 20 initially registered participants, 13 groups sent results.

from the input –this would achieve a precision of 0.198. Secondly, the *Oracle* is the aggregated precision value of words that were correctly guessed by at least one of the participants. With a precision of 0.927, only 7.3% of the OOV words were missed by all of the participants.

The system presented by RAE clearly outperformed the rest of the systems, with 18.7% gain over the runner-up, Citius-Imaxin. Most of the other systems achieved intermediate precision values that range from 54% to 67%. We believe that one of the features that stand out from the winners' systems is the careful integration of different components that consider a number of misspelling cases, as well as the quality and coverage of the components utilized. We comment on the results in detail in Section 6.7.2.

Appendix I shows the list of OOVs that none of the systems has normalized correctly (39 words, 7.25% of the total). This list features a diverse variety of deformations and modifications: for example, the pair *filosofia/Filosofía* (Philosophy) requires correcting capitalization and accents at the same time; the pair *yaa/allá* (there), although not a standard abbreviation, is orthographically very distant and the word *ya* (already) looks like a much more suitable alternative.

6.7 Analysis of Results and Discussion

In this Section, we analyze word categories, the components of the systems, and the techniques and resources they used.

6.7.1 Results by Word Category

Now we delve into the results by breaking down their performance by word category. This allows us to perform a deeper study of the systems' outputs, finding out the categories over which each system performs better.

First of all, all the OOVs in both the development and test corpora were manually categorized into one of the following word categories: acronym, common word, smiley, entity, foreign word, neologism, not in Spanish (NoEs), onomatopoeia, or unsure. Note that *NoEs* words only include those that are part of a tweet written in Spanish, but for some reason it contains some non-Spanish words; *Unsure* was reserved for cases where either the type was unclear, or none of the predefined categories was suitable. Table 6.3 shows the distribution of these categories in the corpora. It can

Rank	System	Prec1	Prec2
—	*Oracle*	0.927	—
1	RAE	0.781	—
2	Citius-Imaxin	0.663	0.662
3	UPC	0.653	—
4	Elhuyar	0.636	0.634
5	EHU	0.619	0.609
6	Vicomtech	0.606	—
7	UArizona	0.604	—
8	UPF	0.548	0.491
9	UAlicante	0.545	0.521
10	UMelbourne	0.539	0.517
11	USevilla	0.396	—
12	UJaen	0.376	—
13	UCoruña	0.335	—
—	*Baseline*	0.198	—

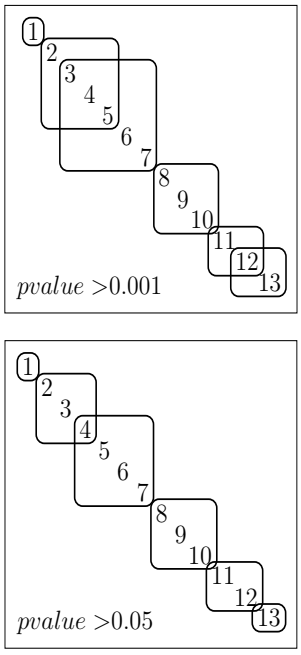
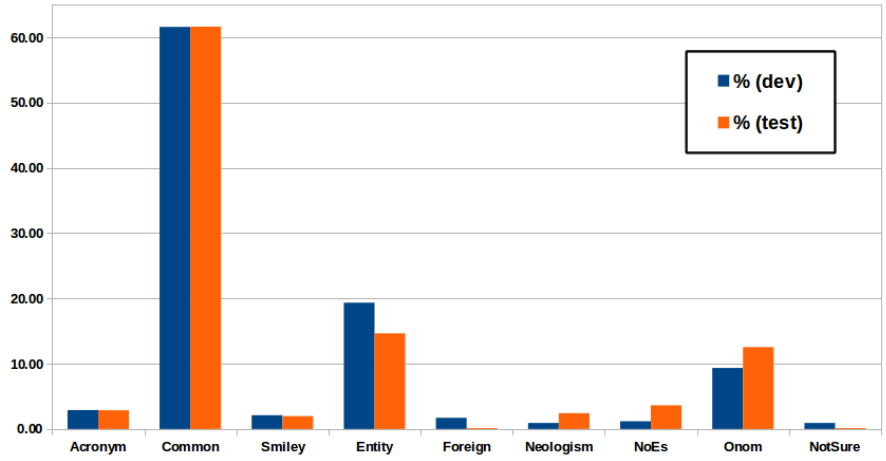


Table 6.2: Precision of the Tweet-Norm 2013 participants. The graphs on the right side show the results of a statistical significance test using McNemar's test. Two systems (based on Prec1) share a cluster if they are not significantly different under the reported *pvalue* (either $0,001$ or $0,05$).

be seen that word categories are similarly distributed in the development and test corpora. In both cases, common words, entities, and onomatopoeias are the most frequent word categories, in that order. The other word categories are less popular, and do not even account for 4% of occurrences in any case.

So how did participants do with respect to each word category? Next, we look at the performance of each participating system broken down into word categories. Table 6.4 shows the precision values by category for the best run for each of the 13 participants, where the rows represent participants and are ordered by overall performance, and the columns represent word categories and are ordered by frequency in the test corpus.

RAE outperformed all the other participants for the three most popular word categories. These are common words, entities, and onomatopoeias, which account for 88.8% of the words in the corpus. The performance gains over the runner-ups for



	Acron.	Common	Smiley	Entity	Foreign	Neolog.	NoEs	Onom.	Unsure
Dev	2.87	61.63	1.96	14.65	0.15	2.42	3.63	12.54	0.15
Test	2.89	61.58	2.11	19.34	1.71	0.92	1.18	9.34	0.92

Table 6.3: Distributions (in percents) of word categories in the development and test corpora.

common words, entities, and onomatopoeias, are 21.9%, 19.2% and 8% respectively, which leads to a large extent to the performance gain of 17.8% over the runner-up for the overall performance, *Citius-Imaxin*.

RAE's system was outperformed by at least another system for some word categories of lower frequency. This includes (i) acronyms, where USevilla performed 50% better, (ii) neologisms, where UArizona performed 30% better, (iii) smileys, where as many as 7 systems performed 44.4% better, and (iv) NoEs, where UArizona and UMelbourne's systems performed 11.1% better. Despite the lower frequency of these word categories, which have little impact on the overall performance, this posits several ways for improving RAE's system. Still, RAE performed better than the average for all word categories.

The most frequent word category (common) accounts for 62% of the OOVs, which is therefore the main factor that determines the final performance of each system. It can be seen that there is a strong correlation between the ranking based on the overall performance, and the ranking based on the performance for common words. There are just a few exceptions at the bottom of the ranking, but the ranking would be the same for the top 6 systems if we only considered common words. For some of

the exceptions at the bottom of the ranking, such as UAlicante, USevilla, UJaen, and UCoruña, the performance for entities is substantially lower than their performance for common words and onomatopoeias (as low as 1% in the case of UJaen). The fact that these systems performed very differently for the other frequent categories, i.e., entities and onomatopoeias, explains this difference between the common and total scores. The performance drop for the entity category may occur in part due to the use of their own OOV detection mechanism, as explained in Section 6.7.3.

The oracle shows the performance of the best possible system by considering all the words that were accurately normalized by at least one participant. With a 0.927 overall performance for the oracle, it shows that only 7.3% of the words were not normalized correctly by any of the participants. However, while common words represent the most frequent category in the corpus, it is also the one with highest percentage of words missed by everyone (15.8%). The second most frequent category, entities, had also 8.1% of the words missed by everyone. This posits the need for further exploring normalization of these two categories, common words and entities, in future research. Improving these would substantially improve the overall performance of normalization systems. On the other hand, a look at the average performance shows that acronyms were the most difficult overall, and certainly a word category that needs careful analysis for improvement in future work.

6.7.2 Focused phenomena

The good performance of the RAE system is remarkable. It for the most part outperforms all the others with a 78% score for precision, while most systems score between 54% and 67%. The difference can be explained by the thorough and detailed treatment of many linguistic phenomena appearing in Twitter, the statistical combination of the used modules, and the quality and coverage of used resources. Note that Han et al. (2013) describe a system for English that achieves 72.3% F-score in a similar scenario to that described in our proposal: OOV detection + normalization. Only the first Spanish system in the competition outperforms this score. The main diffence with regard to our evaluation protocol is that we use one-to-many correction pairs, while Han et al. (2013) only use one-to-one pairs.

The phenomena explicitly addressed by several of the participant systems are the following:

- Usual orthographic mistakes ($h \rightarrow$).
- Usual phonological changes ($c/qu \rightarrow k$).

Rank	System	Com.	Ent.	Ono.	NoEs	Acr.	Neol.	Smil.	Fore.	Unsu.	Total
—	*Oracle*	0.842	0.919	1.000	0.938	1.000	0.870	1.000	1.000	0.952	0.927
1	RAE	**0.806**	**0.897**	**0.651**	0.750	0.421	0.625	0.692	**1.000**	**1.000**	0.781
2	Citius-Imaxin	0.662	0.753	0.554	0.750	0.474	0.563	**1.000**	**1.000**	0.000	0.663
3	UPC	0.652	0.701	0.602	0.667	0.368	0.625	**1.000**	**1.000**	**1.000**	0.653
4	Elhuyar	0.630	0.711	0.542	0.750	0.526	0.438	**1.000**	**1.000**	**1.000**	0.636
5	EHU	0.625	0.649	0.530	0.667	0.368	0.688	**1.000**	**1.000**	0.000	0.619
6	Vicomtech	0.610	0.670	0.530	0.417	0.579	0.438	**1.000**	**1.000**	**1.000**	0.606
7	UArizona	0.588	0.598	0.530	**0.833**	0.579	**0.813**	**1.000**	**1.000**	0.000	0.604
8	UPF	0.576	0.649	0.578	0.292	0.211	0.250	0.154	0.000	0.000	0.548
9	UAlicante	0.598	0.433	0.590	0.292	0.421	0.438	0.154	**1.000**	**1.000**	0.545
10	UMelbourne	0.471	0.732	0.434	**0.833**	0.526	0.750	**1.000**	**1.000**	**1.000**	0.538
11	USevilla	0.407	0.258	0.349	0.417	**0.632**	0.375	0.923	**1.000**	**1.000**	0.396
12	UJaen	0.502	0.010	0.494	0.000	0.000	0.125	0.000	0.000	0.000	0.376
13	UCoruña	0.456	0.041	0.373	0.042	0.000	0.000	0.000	0.000	0.000	0.335
—	*Baseline*	0.526	0.032	1.000	0.557	1.000	0.750	0.917	1.000	0.049	0.196
—	*Average*	0.583	0.546	0.520	0.516	0.393	0.471	0.686	0.769	0.538	0.562
—	*Best*	0.806	0.897	0.651	0.833	0.632	0.813	1.000	1.000	1.000	0.781

Table 6.4: Precision values broken down into word categories for the best run for each of the participants, and average and best performance for each word category. Participants in rows are ordered by overall performance, while word categories in columns are ordered by their frequency in the test corpus. Bold figures represent participants who obtained the highest precision score for the word category in question.

- Omission of graphical accent (*á* → *a*).
- Omission of characters, mainly vowels and final letters, especially in participles (*encantado* → *encantao*).
- Use of abbreviations or reduction of words to their initial characters (*examen* → *exam*)
- Emphasis expressed via character repetition (usually vowels) (*felicidades* → *felicidadeeees*).
- Omitted capitalization (*Juan* → *juan*).
- Contiguous word joining *es que* → *esque*.
- Logograms and pictograms. (*por* → *x*; *dos* → *2*).
- Repetition of onomatopoeias (*ja* → *jajajaja*).

The lexica used by the participants to look for normalized variations are mostly Spanish dictionaries, spell checkers, and also Freeling —i.e., the same tool used for the preprocessing step. Some have also used other resources: (i) English dictionaries to look for OOV words that, without being Spanish words, do not need to be changed, (ii) the Spanish Wikipedia[13] to identify named entities, (iii) small slang and variation dictionaries, and (iv) word frequencies extracted from other corpora to identify common misspellings on the Internet and Twitter.

Different approaches that make use of language models have also relied on several corpora of Spanish language texts. Both general purpose corpora and specific Twitter corpora have been used to create language models. One of the participating systems used the API of a search engine to filter multi-word terms.

The participants also utilized several tools to create their normalization systems. Many used spell checkers (e.g., Aspell,[14] Hunspell,[15] Jazzy[16]), which can also be used to look for alternative candidates. Some also used Foma[17] to work with transducers, which learns transformation rules for phonemes and graphemes. In some cases, transformation rules have also been defined based on language models, e.g., using Phonetisaurus.[18] For the selection of the final candidate, some relied

[13] http://es.wikipedia.org
[14] http://aspell.net
[15] http://hunspell.sourceforge.net
[16] http://jazzy.sourceforge.net
[17] https://code.google.com/p/foma/
[18] http://code.google.com/p/phonetisaurus/

Rank	System	Architecture	Filtering	LM	Rules	Ap. Search	SMS	Phonetics	NEs
1	RAE	G/F	LM	3-gram	R	ED	–	PH	E
2	Citius-Imaxin	G/F	LM	2-gram	R	ED	–	–	E
3	UPC	G/F	S(vot)	–	R	ED	–	PH	E
4	Elhuyar	G/F	LM	3-gram	R	LCS	SMS	–	E
5	EHU	PP	–	1/2-gram	R	–	–	–	E
6	Vicomtech	G/F	LM+S	5-gram	–	ED	–	–	E
7	UArizona	PP	–	–	R	–	–	–	–
8	UPF	PP	–	–	–	Spell	SMS	–	I
9	UAlicante†	G/F	LM	3-gram	R	LCS	–	MPH	I
10	UMelbourne	G/F	S(dist)	–	–	ED	SMS	MPH	I
11	USevilla†	G/F	S(conf)	–	R	ED	–	–	–
12	UJaen	G/F	F	–	R	Spell	SMS	–	E
13	UCoruña†	G/F	S	–	R	Spell	SMS	–	–

Table 6.5: Synoptic table of system's charecteristics. See Section 6.7.3 for details.

on corpora-based frequencies, whereas others used language modeling tools (e.g., OpenGrm[19] and SRILM[20]).

6.7.3 Summary of Techniques and Resources

Table 6.5 summarizes the characteristics of each system participating in the Tweet-Norm 2013 evaluation. These characteristics include only the best run from each participant. The table is divided into four parts according to the clusters obtained in Table 6.2 for a *pvalue* > 0.001: 1, 2–7, 8–10, 11–13. There are eight columns containing the analyzed characteristics for each system. Their meaning is as follows:

1. Architecture: **G/F**: Generate/Filter architecture. A generation process proposes a set of alternative spellings (IV words) for each OOV, this is called the *confusion set*. In a second step a filtering mechanism is implemented to select one of the proposed words in the confusion set. This architecture is used by 10 out of the 13 systems. **PP**: Pipeline architecture. Each OOV word goes through a sequence of analyzers. The process stops when an IV word is produced by one of the analyzers.

2. Filtering Mechanisms (for G/F architectures): **LM**: a Language Model selects the most probable candidate from the confusion set according to the context

[19] `http://www.opengrm.org`
[20] `http://www.speech.sri.com/projects/srilm/`

words. **S**: some Scoring function is derived from the generation of IV words. This may implement voting or other confidence estimation techniques such as distributional similarity. **F**: The most frequent word in a sample corpus is selected. It is a particular case of LM of length 1.

3. Language Model: n**-gram**: A language model of length n is used (not necessary as a Filtering mechanism).

4. Rules: **R**: Some kind of knowledge-based transformation rules are implemented (e.g., shorthands, phonographemes). The actual number and complexity of the rules may be disparate.

5. Approximated Search: **ED**: Edit Distance is used to find similar IV words. **LCS**: Longest Common Subsequence is used to find similar IV words. **Spell**: A spell-checker is used to find similar IV words.

6. SMS: **SMS**: The system uses dictionaries of textese utilized in SMS, and slang.

7. Phonetics: **MPH**: Phonetic representations of words are obtained with the Metaphone algorithm. **PH**: Words are represented with IPA phonemes.

8. Named Entities: **E**: Explicit lists of NEs are compiled from one or more sources as new IV words. **I**: The dictionary of IV words is enriched with textual sources thus, NEs are implicitly added.

A dash (i.e. –) is used for the systems that do not have that feature. Finally, the dagger (i.e. †) notes that the system uses its own OOV detection mechanism instead of the ones provided in the test set. In this case, we have to be cautious when drawing comparisons with the rest of the systems.

6.7.4 Discussion

According to the clusters obtained in Table 6.2 for a $pvalue > 0.001$, we can divide the systems in four groups in this way: 1, 2–7, 8–10, 11–13. Looking at the columns of Table 6.5, we can characterize the systems within each group as follows:

- Five out of the top seven systems use a generate/filter architecture with a language model as filter, while in the lower half of the table many systems use local or confidence based scoring mechanisms. This suggests that a filter operating on a OOV's confusion set cannot work solely on the intrinsic properties of the words but on its context.

- Using a language model seems to be a competitive way of scoring and selecting a good normalization in the generate/filter architecture, and it is consistently better than other scoring methods devised by the participants.
- The top six systems have compiled extensive lists of named entities, while the rest have not targeted this kind of knowledge or have done so indirectly. Surprisingly, the precision for entity normalization does not seem to be explained by this feature, neither by the use of SMS lexica.
- Using the phonetic representation of words seems to be positive although not as much when the Metaphone algorithm is used. This suggests that Metaphone may not be a suitable method for generating alternatives. Unfortunately, the performance of the two systems using Metaphone (UAlicante, UMelbourne) varies greatly in the three most frequent categories.
- One reason that explains the good results of the RAE system is that, compared to others, their mechanism for confusion set generation seems very precise (Porta and Sancho, 2013) and very little noise has to be filtered out afterwards. For other systems that, like UPC, can generate thousands of alternative spellings (Ageno et al., 2013), it becomes much more difficult to select the correct candidate. EHU and UArizona rely mainly on a sound set of transformation rules (hand-crafted for UArizona) and also achieve good results with a pipeline architecture.
- UMelbourne is especially good in the NoEs and neologism categories. Their approach based on distributional semantics on a large corpus of tweets may explain their good results. Under the hypothesis that the use of neologisms and foreign words in Twitter is some kind of slang (i.e., slang replaces well-known words, it is informal, the users are familiar with its context of application), the vocabulary size of these categories is probably much reduced when compared to common words. Therefore, their context of usage may be more accurately characterized by distributional semantics. This reasoning can be applied to some extent to the entity category, in which UMelbourne has notable results too.

6.8 Conclusions and future work

The development of the benchmark evaluation framework and the *TweetNorm_es* corpus, as well as the Tweet-Norm 2013 shared task that enabled evaluation

of systems from 13 participants, served as an initial step toward encouraging implementation of new methods for and approaches to Spanish microtext normalization in the research community. The high number of participants has proven the task relevant, and posited a number of issues to be considered in future research.

The work presented in this paper paves the way for future development and research on Spanish microtext normalization, setting forth a methodology to create a corpus for these purposes, as well as releasing the corpus we created following such methodology. The corpus provides a gold-standard for development and evaluation of microtext normalization tools.

The corpus is available under the terms of the CC-BY license for anyone interested in the task, and can be found at the website of the workshop.[21]

This work has also brought to light a number of issues that remain unresolved and are worth studying in future work. Here we have performed *in vitro* evaluations of the normalization systems. We believe that *in vivo* evaluations by incorporating normalization into other NLP systems, such as sentiment analysis or machine translation will enable deeper study of the task, as well as to quantify the actual effect of processing normalized outputs. Additionally, we would like to broaden the task by not only dealing with lexical normalization, but also addressing complementary tasks such as normalization of syntax and/or real-word errors. Last but not least, we are aware that the size of the corpus is limited. Extending the corpus and considering different OOV categories would enable to perform a more detailed evaluation, especially for machine learning purposes.

Acknowledgments

We would like to thank all the members of the organizing committee. This work has been supported by the following projects: Spanish MICINN projects *Tacardi* (Grant No. TIN2012-38523-C02-01), *Skater* (Grant No. TIN2012-38584-C06-01), *TextMESS2* (TIN2009-13391-C04-01), *OntoPedia* (Grant No. FFI2010-14986) and *Holopedia* (TIN2010-21128-C02-01); *Xlike* FP7 project (Grant No. FP7-ICT-2011.4.2-288342); UNED project (2012V/PUNED/0004); *ENEUS-Marie Curie Actions* (FP7/ 2012-2014 under REA grant agreement nº302038); *Celtic* CDTI FEDER-INNTER-CONECTA project (Grant No. ITC-20113031); Research Network MA2VICMR (S-2009 / TIC-1542); and *HPCPLN* (Grant No. EM13/041, Xunta de Galicia).

[21] `http://komunitatea.elhuyar.org/tweet-norm/`

Appendix I: List of unresolved OOV words

Table 6.6 contains the list of words from the corpus that none of the systems found the correct variation for. The list comprises the word as spelled originally in the corpus on the left column, and the correct variation annotated manually on the right column.

Original	Variation
FYQ	Física_y_química
sisiii	sí_sí
yaa	allá
picolos	picoletos
nainonainonahh	nainonainoná
gordys	gorditas
JUUUM	hum
Tuitutil	TuitÚtil
crst	Cristo
mencantaba	me_encantaba
diitaas	diítas
soo	eso
queeee	qué
Teinfiniteamo	Te_amo_infinitamente
aber	a_ver
Hum	Humedad
L.	l.
Muchomuchacho	Mucho_Muchacho
Hojo	Jo
jonaticas	jonáticas
gafis	gafitas
her	hermano|hermana
MIAMOR	mi_amor
guapii	guapita
WAPAHHH	guapa
EAEA	ea_ea
Acho	Macho
tirantitas	tirantitos
HMYV	MHYV
filosofia	Filosofía
nah	nada
FAV	favorito
JIIIIIIIIOLE	Olé
Fotazo	fotaza
gor	gorda|gordo
coner	con_el
shh	sí|sé
primera+	primera_más
salobreja	Salobreja

Table 6.6: List of OOV words for which none of the participants found the correct variation.

CHAPTER 7

Microtext Normalization System

Elhuyar at TweetNorm 2013

Xabier Saralegi, Iñaki San Vicente

Elhuyar Fundazioa

This paper presents the system developed by Elhuyar for the Tweet-Norm evaluation campaign which consists of normalizing Spanish tweets to standard language. The normalization covers only the correction of certain Out Of Vocabulary (OOV) words, previously identified by the organizers. The developed system follows a two step strategy. First, candidates for each OOV word are generated by means of various methods dealing with the different error-sources: extension of usual abbreviations, correction of colloquial forms, correction of replication of characters, normalization of interjections, and correction of spelling errors by means of edit-distance metrics. Next, the correct candidates are selected using a language model trained on correct Spanish text corpora. The system obtained a 68.3% accuracy on the development set, and 63.36% on the test set, being the 4th ranked system on the evaluation campaign.

Published in Proceedings of *"XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural"*. Tweet Normalization Workshop (Tweet-Norm at SEPLN 2013) Madrid, 2013. pp. 64-68. ISBN: 978-84-695-8349-4.

7.1 Introduction

Social media and specially Twitter have become a valuable asset for information extraction purposes. Twitter falls into the category of "microtext". As such, tweets present some characteristics which limit the straight application of natural language processing techniques: non standard orthography, colloquial expressions, abbreviations... So, converting Twitter messages to standard language is an essential step before applying any linguistic processing.

This paper presents the system developed by Elhuyar for the TweetNorm task, a task which consists of normalizing Spanish tweets. The normalization just covers the correction of certain OOV words. After tagging the reference using FreeLing (Padró et al., 2010), those words without analysis are regarded as OOV. The OOV list was provided by the organizers. Real-word errors are not treated in this task, that is, cases where a word is misspelled but the misspelled form also exists in the dictionary (e.g., *'té' -tea-* and *'te' -to you-*).

The developed system follows a two step strategy. First, candidates for each problematic word are generated by means of various methods dealing with the different error-sources: extension of usual abbreviations, correction of colloquial forms, correction of replication of characters, normalization of interjections, and correction of orthographical errors by means of edit-distance metrics. The second step selects the correct candidate, by comparing the adequacy of each candidate against a language model trained from standard Spanish text corpora. The EFE news corpus and the Spanish Wikipedia were used for such purposes. The system obtained a 68.3% accuracy on the development set, and 63.6% on the test set, being the 4th ranked system on the evaluation campaign.

7.2 Related Work

In the last few years many researchers have turned their efforts to microblogging sites such as Twitter. However, the special characteristics of the language of Twitter require a special treatment when analyzing the messages. A special syntax (RT, @user, #tag,...), emoticons, ungrammatical sentences, vocabulary variations and other phenomena lead to a drop in the performance of traditional NLP tools (Foster et al., 2011; Liu et al., 2011).

To solve this problem, many authors have proposed a normalization of the text, as a pre-process of any analysis, reporting an improvement in the results. Han

and Baldwin (2011) use morphophonemic similarity to match variations with their standard vocabulary words, although only 1:1 equivalences are treated, e.g., *'imo = in my opinion'* would not be identified. Instead, they use an Internet slang dictionary to translate some of those expressions and acronyms. Liu et al. (2012) propose combining three strategies, including letter transformation, "priming" effect, and misspelling corrections.

7.3 Our System

The system performs the normalization process of tweets in two steps (see Figure 7.1). In a first step several methods are applied for generating candidates for the OOV words. In the next step a single candidate is selected for each OOV word by using language models.

Two data-sets were provided by the organizers of the Tweet-Norm event. One development-set C_{dev} composed of 500 tweets, and one test-set C_{test} composed of 600 tweets which was used only for evaluation purposes.

7.3.1 Generation of candidates

Some of these methods use reference lexicons for generating candidates. A reference lexicon of correct forms D_r was built by joining the FreeLing's dictionary forms and forms extracted from the EFE news corpus (146M words) and Spanish Wikipedia corpus (41M words), which theoretically include correctly written texts. A minimum frequency threshold was established in order to avoid possible typos, because several of them were found in both EFE and Wikipedia (e.g., *'tambien'*). A disadvantage of using these corpora is that they are focused on formal registers while the register of twitter is more informal. However, it is a difficult task to compile a corpus for informal register without including many wrongly written texts. So we sacrificed register adaption in benefit of correctness.

Colloquial vocabulary (COL)

We created a list of colloquial vocabulary (e.g., *'colegui', 'caseto', 'bastorro'*) by collecting words from two sources: *"Diccionario de jerga y expresiones coloquiales"*[1] dictionary and *www.diccionariojerga.com*, a crowdsourcing web including colloquial

[1] http://www.ual.es/EQUAL-ARENA/Documentos/coloquio.pdf

vocabulary edited by users. A different word corresponding to the correct form was inserted if necessary, otherwise the word itself is inserted as correct form. This list $L_c = \{(c_i, c'_i)\}$ contains 1088 entries.

The method based on this list is simple, if an OOV word c_i is included in the list the corresponding correct form c'_i is generated as a candidate.

Abbreviations (ABV)

A list containing the most used abbreviations (e.g., *'mñn'→'mañana'*) and contractions (forms that join more than one word, e.g., *'porfa'→'por_favor'*) in Twitter was created. First, the most frequent OOV words of a Twitter corpus (309,276 tweets, 4M tokens) were extracted, and the top 1,500 candidates ($freq(abv_i) > 25$) were analyzed, looking for abbreviations and contractions. Their corresponding correct forms were established by hand. As a result, 188 abbreviations were included in the list $L_{abv} = \{(abv_i, abv'_i)\}$. As with the previous method, for each OOV abv_i included in the list its standard form abv'_i is proposed as a correct candidate.

Interjections (INTJ)

Regular expressions were created for matching and normalizing the most common interjections and their variations (e.g., *'jeje', 'puf'*), identified in the development corpus C_{dev}.

Repeated letters (REP)

Repeated letters are removed from an OOV word if it does not appear in the reference lexicon D_r. Then if the modified form appears in D_r (e.g., *'calooor'→'calor'*) it is included as candidate.

Proper Nouns (PN)

A list of usual proper nouns was built from the Wikipedia corpus. Words in uppercase w_{uc} with a minimum frequency ($freq(w_{uc}) > 100$) and whose frequency is higher than that of their form in lowercase ($freq(w_{uc}) > freq(w_{lc})$) are taken as secure proper nouns. 6,492 words were collected in this manner.

If an OOV word w appears in a list of usual proper nouns and its first character is in lowercase then it is put in uppercase (e.g., *'betis'→'Betis'*).

Uppercase (UC)

If all characters of an OOV w_{uc} word are in uppercase the following rules are applied:

- If w_{uc} appears as it is in D_r, w_{uc} is proposed as candidate (e.g., *'IVA'→'IVA'*).
- If w_{uc} is included in D_r in lowercase $w_{lc} = lc(w_{uc})$, then w_{lc} is proposed as candidate (e.g., *'IMPORTANTE'→'importante'*).
- If w_{uc} is included in D_r with the first character in uppercase $w'_{uc} = ucfirst(w_{uc})$, then w'_{uc} is proposed as candidate (e.g., *'MADRID'→'Madrid'*).

Spelling errors (COG)

String similarity measures are useful for detecting correct forms of misspelled words. If the string similarity between an OOV word w and a correctly written word w' exceeds a certain threshold we can take w' as a correct candidate. We apply edit distance as follows: first, a set of transliteration rules are applied to both words ($trans(w)$ and $trans(w')$) in order to normalize some characters (e.g., $b = v$, $ki = qui$, $ke = que$ ...). Then, Longest Common Subsequence Ratio (LCSR) is calculated between $trans(w)$ and $trans(w')$. In order to reduce the computational cost of the process, LCSR is only computed for those words in our lexicon D_r that share the first character (except for h) with w and have a similar length ($\pm 20\%$). LCSR gives a score between 0 (minimum similarity) and 1 (maximum similarity). Those forms that reach a score greater than 0.84 are taken as candidates.

7.3.2 Selection of correct candidates

A tweet $t = \{f_0, ..., f_i, ...f_n\}$ can contain more than one OOV word, and each OOV word f_i can have several candidates $\{f_{i0}, ..., f_{ij}, ..., f_{im}\}$ after applying the above-mentioned methods (see Figure 7.1). Thus, a disambiguation process must be applied in order to obtain a single correct candidate for each OOV word. For that aim we use language models. The system selects for each tweet, the combination of candidates that best fits the language model, that is, the combination which maximizes the log probability of the sequence of words.

SRILM toolkit (Stolcke, 2002) was used for training and applying the language model. For the training process the EFE news corpus and the Spanish Wikipedia corpus were used. As mentioned in section 7.3.1, we chose those sources in order to guarantee maximal language correctness.

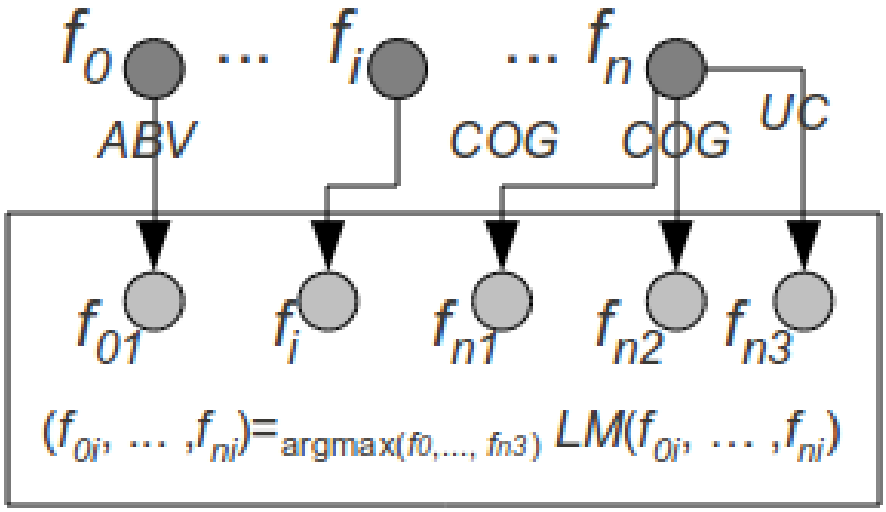


Figure 7.1: The diagram shows the two steps of the normalization process.

7.4 Results

Table 7.1 shows the results for the experiments done on the 500 tweets of the development collection C_{dev}, depending on the different treatments and disambiguating by using an unigram language model trained on Wikipedia and EFE corpora. The baseline consists of selecting the OOV itself as correct candidate.

All the methods proposed provide an improvement over the baseline except for the UC method (see Table 7.1). The degree of improvement provided by each method varies depending on the frequency of the error-type treated by the method and the performance of the method itself. Thus, the string similarity based method COG provides the highest improvement (76.93% over the baseline), which means that the presence of typos is high and that the performance of the method is good. Both REP and ABV methods offer an improvement around 40% over the baseline. The treatment of interjections (INTJ) is also important, providing an improvement of 25% over the baseline. The error-types treated by the COL and UC are very scarce (14 and 5 respectively on C_{dev},). In the first case, although the methods perform well, the improvement is small. In the case of UC, most of the cases (4 out of 5) concatenate various error-types, and our system can not deal with error concatenations, leading to a performance decrease. Nevertheless, the method does provide an improvement

when it is used combined with a bigram or a trigram LM, and thus, we include it in the all configuration. Error-types treated by PN are a bit more frequent ($\simeq 40$ in C_{dev}). Although the method is quite precise ($P \simeq 80\%$) it lacks coverage ($R \simeq 60\%$). Among all method combinations the best accuracy was achieved when all of them were combined (ALL). So, we conclude that the LM manages properly the candidates provided by all the methods.

	Acc. on the Devel. set	**Improvement over Baseline**
Baseline	23.28	-
Baseline+COL	24.2	3.95%
Baseline+ABV	32.16	38.14%
Baseline+INTJ	29.1	25%
Baseline+REP	34	46.05%
Baseline+PN	24.81	6.57%
Baseline+UC	23.12	-0.69%
Baseline+COG	41.19	76.93%
ALL	66.16	184.19%

Table 7.1: Accuracies for the candidate generation methods. Last column shows the improvement the method achieves over the baseline.

We performed further experiments with different orders of n-grams and different configurations of corpora, using in all cases the ALL configuration. According to the results (table 7.2), when larger orders of n-grams are used higher accuracies are obtained. This improvement is significant between 1-gram and 2-gram models. There is no improvement when using larger orders of n-grams. As for the corpora used, combining Wikipedia and EFE corpora provides the best performance. So it seems that they complement each other. Thus, evaluation over the test-set C_{text} was carried out using the bigram LM trained over the joint corpus between EFE and Wikipedia (See fifth column in Table 7.2).

Error analysis

We performed error analysis over the OOV words not treated correctly by our best system for the 500 tweets of the development collection C_{dev}. Following, we explain the main problems detected in our system:

- Concatenation of errors: Generation methods are not combined between each

	Development			Test		
	unigr.	**bigr.**	**trigr.**	**unigr.**	**bigr.**	**trigr.**
EFE	64.93	66.62	66.62	-	-	-
Wikipedia	65.54	67.69	67.69	-	-	-
EFE + Wikipedia	66.16	**68.30**	67.69	-	63.60	-

Table 7.2: Accuracies for the different language models' experiments, using the ALL configuration for the generation of candidates.

other because LM is not capable of properly managing the noise created (e.g., *'SOI'→'SOY'→'soy', 'cumpleee'→'cumple'→'cumpleaños'*).

- Abbreviations and contractions: The abbreviation and contractions not included in our list are not properly normalized (e.g.,*'cmun'→'común', 'deacuerdo'→'de_acuerdo'*). LCSR based method is not capable of finding the correct form for the case of abbreviations either, because the distance is very high. If the threshold is decreased too much noise is created.
- Lack of domain adaptation: LM is trained from corpora corresponding to news and Wikipedia domains where informal register is not included. Because of that there are some colloquial expressions (e.g., *'maricón', 'bonico', 'comidita'*) and proper nouns (e.g., *'Pedrete', 'Fanegol'*) that are not included in our reference lexicon D_r and which are not properly disambiguated.
- Keyboard typos: Some errors correspond to key confusion at writing time. In some cases LCSR is not reached. (e.g., *'pa'→'la', 'tenho'→'tengo'*).

7.5 Conclusions

This paper presents a system for normalizing tweets written in Spanish. The system first generates a number of possible correction candidates for OOV words and then selects the candidate that better matches a language model trained over corpora of standard Spanish. Our system achieved the 4th rank among thirteen contestants in the tweet-Norm evaluation campaign. We consider this a satisfactory performance taking into account that, aside from the best system, the next four contestants are quite close to each other. Furthermore, our error analysis has shown that we still have room for improvement.

Edit distance must be adapted to better deal with abbreviations, contractions and keyboard errors. An alternative to improve that aspect could be to use a more complex strategy based on finite state toolkits such as Foma (Hulden, 2009).

On the other hand, we apply the different candidate generation methods in parallel, they are not combined in any way. This leads to a poor performance when an OOV has several errors concatenated. Therefore, we should explore possible method combinations, avoiding at the same time to generate too much noise, because the LMs would lose disambiguation capacity. In addition, we could experiment with larger LMs, and also LMs that are more focused on informal register.

Acknowledgments

This work has been partially founded by the Industry Department of the Basque Government under grant IE11-305 (KnowTOUR project).

PART III

POLARITY CLASSIFICATION

III Polarity Classification

This part covers the work done on polarity classification. The papers presented in the following chapters focus on developing polarity classifiers that were tested against well established benchmark datasets for Spanish and English. Since then, new annotated tweet datasets have been created for Basque, Spanish, English and French, as well as polarity models trained on those datasets. A detailed list of generated resources is given in section 11.3.2

Chapter 8 (San Vicente and Saralegi, 2014) summarizes our research on Spanish tweet polarity classification. This research started by implementing a supervised SVM classifier (Saralegi and San Vicente, 2012) for the TASS 2012 evaluation campaign (Villena-Román et al., 2012). That first system was improved by adding new features (Saralegi and San Vicente, 2013) and preprocessing steps derived from the findings in microtext normalization (see part II). The classifiers developed obtained the best results for both TASS 2012 and 2013 Villena-Román et al. (2014) campaigns. Furthermore, domain adaptation experiments were carried out for the tourism domain in (San Vicente and Saralegi, 2013). Lastly, (San Vicente and Saralegi, 2014) added new lexicon and syntax-based ngram features, providing the second best system in the TASS 2014 task (Román et al., 2015).

Chapter 9 (San Vicente et al., 2015) describes how the system developed for Spanish was ported into English, showing competitive results in the ABSA shared task (Pontiki et al., 2015). The paper presents two algorithms for the tasks of opinion target extraction (OTE) and polarity classification respectively. OTE is addressed by means of an averaged Perceptron sequence labeller (Agerri and Rigau, 2016). (San Vicente et al., 2015) sets a milestone, because it provides the first implementation of the Sentiment Analysis tool EliXa, including multilingual capabilities and microtext normalization. EliXa has been further developed to include Basque and French Sentiment Analysis since then.

The publications included in this part (by order of appearance) are listed below. Furthermore, for each publication we state the contribution of the author of this thesis.

- Iñaki San Vicente and Xabier Saralegi. Looking for features for supervised tweet polarity classification. In *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS2014)*, Girona, Spain, September 2014

  Contribution to the paper: Main author of the paper. Responsible for the coding and data processing. Both authors contributed equally in the design of the experiments and manual evaluation, as well as in the writing of the paper.

- Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. Elixa: A modular and flexible absa platform. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 748–752, 2015

  Contribution to the paper: Main author was responsible of the implementation of the software and carried out the experiments related to polarity classification. Second author took part in the design of the experiments and contributed to the writing of the paper. Rodrigo Agerri was responsible for the opinion target extraction experiments and submission.

CHAPTER 8

Spanish Polarity Classification

Looking for Features for Supervised Tweet Polarity Classification

Iñaki San Vicente, Xabier Saralegi

Elhuyar Fundazioa

This article describes the system presented by Elhuyar for the task 1 of the TASS 2014 sentiment analysis evaluation campaign. Our system implements a Support Vector Machine (SVM) algorithm. The system combines the information extracted from polarity lexicons with linguistic features. Incorporating syntax based ngrams and enriching the polarity lexicons prove to be the most influential factors in the improvement of the system with respect to our TASS 2013 participation. The system achieves an 61% accuracy fine granularity and an 69% accuracy for coarse granularity polarity detection.

Published in *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS2014)*, Girona, September 2014.

8.1 Introduction

Knowledge management is an emerging research field that is very useful for improving productivity in different activities. Knowledge discovery, for example, is proving very useful for tasks such as decision making and market analysis. With the explosion of Web 2.0, the Internet has become a very rich source of user-generated information, and research areas such as opinion mining or sentiment analysis have attracted many researchers. Prove of that is that in the last years a growing number of Sentiment Analysis related shared tasks have been organized, such as TASS workshops (Villena-Román et al., 2012; Villena-Román et al., 2014), SemEval shared tasks (Nakov et al., 2013; Pontiki et al., 2014; Rosenthal et al., 2014) or the Concept-Level Sentiment Analysis Challenge at ESWC2014[1].

Being able to identify and extract the opinions of users about topics, events, or products is becoming an essential part of market analysis and reputation management systems, and social media is the main source for such information. Because of its special nature (limited length, non standard language), extracting such information from Twitter presents a challenge for Natural Language Processing systems. The TASS evaluation workshop aims "to provide a benchmark forum for comparing the latest approaches in this field". Our team only took part in the first task, which involved predicting the polarity of a number of tweets, with respect to 6-category classification, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. It must be noted that most works in the literature only classify sentiments as positive or negative, and only in a few papers are neutral and/or objective categories included. We developed a supervised system based on a polarity lexicon and a series of additional linguistic features.

The rest of the paper is organized as follows. Section 8.2 reviews the state of the art in the social media polarity detection field, placing special interest on Twitter and its special characteristics. The third section describes the system we developed, the features we included in our supervised system and the experiments we carried out over the training data. The next section presents the results we obtained over the test data-sets. The last section draws some conclusions and future directions.

[1] `http://challenges.2014.eswc-conferences.org/index.php/SemSA`

8.2 State of the Art

Much work has been done on the sentiment analysis field, from polarity lexicon induction to sentiment labeling and opinion extraction. There are extensive surveys on the field (Pang and Lee, 2008; Liu, 2012). In the last years microblogging sites such as Twitter have attracted the attention of many researchers with diverse objectives: stock market prediction (Bollen et al., 2010), polling estimation (O'Connor et al., 2010) or crisis situations analysis (Nagy and Stamberger, 2012).

The special characteristics of the language of Twitter require a special treatment when analyzing the messages. A special syntax (RT, @user, #tag,...), emoticons, ungrammatical sentences, vocabulary variations and other phenomena lead to a drop in the performance of traditional NLP tools (Foster et al., 2011; Liu et al., 2011). In order to solve this problem, a normalization of the text has been proposed (Brody and Diakopoulos, 2011; Han and Baldwin, 2011), as a preprocess of any analysis.

Once the normalization has been performed, traditional NLP tools may be used to analyze the tweets and extract features such as lemmas or POS tags (Barbosa and Feng, 2010). Emoticons are also good indicators of polarity (O'Connor et al., 2010). Other features analyzed in sentiment analysis such as discourse information (Somasundaran et al., 2009) can also be helpful. Speriosu et al. (2011) explore the possibility of exploiting the Twitter follower graph to improve polarity classification, under the assumption that people influence one another or have shared affinities about topics. Sindhwani and Melville (2008) adopt a semi-supervised approach using a polarity lexicon combined with label propagation. (Barbosa and Feng, 2010; Kouloumpis et al., 2011) combined polarity lexicons with machine learning for labeling sentiment of tweets. We adopt this strategy too, which has proven a successful approach in previous shared tasks (Saralegi and San Vicente, 2012; Mohammad et al., 2013).

8.3 Experiments

8.3.1 Training Data

The same as in previous editions, the training data C_t consists of 7,219 Twitter messages. Each tweet is tagged with its global polarity, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. 6 levels have been defined: two positive (P and P+), two negative (N and N+), neutral

(NEU) and no sentiment (NONE). The corpus is skewed towards positive polarity (see category distribution in the second column of Table 8.4), having nearly the 40% of the tweets P or P+ category.

8.3.2 Polarity Lexicon

Elhuyar Polar

Our main resource is the Elhuyar Polar (ElhPolar) polarity lexicon which was created for previous editions of the TASS workshop. The lexicon was semiautomatically built, on the one hand, by translating an existing English lexicon, and on the other by extracting positive and negative words from the training corpus C_t relying on association measures. All polarities in the lexicon were manually corrected by two annotators, in order to ensure their correctness to the greatest extent. A detailed explanation of building process is included in (Saralegi and San Vicente, 2013). In addition, for TASS 2014 edition, ElhPolar was enriched with a manually compiled list of locutions, mainly verbals ("*agachar las orejas*", "*mantener el tipo*"), and some set phrases ("*ir a por lana y salir trasquilado*").

Additional lexicons

Experiments were conducted in order to include other polarity lexicons. Combining polarity lexicons will allow us to increase the coverage of the lexicon. We want to stress that even if we are trying to improve the coverage of our lexicon, it is important for us to minimize the noise other lexicons may introduce. That is why we gave preference to manually corrected resources and took some measures to discard entries which may have ambiguous (e.g., "infantil") or weak polarities (e.g., "desechable"). Table 8.1 provides statistics of the lexicons used. Following we describe briefly the lexicons used in our experiments:

- *Mihalcea's Lexicon (Perez-Rosas et al., 2012) (Mih)*: Perez Rosa's paper describes two lexicons. We only use here the one regarded as *"full strength"* lexicon, because it integrates manual annotations from OpinionFinder (Wilson et al., 2005).
- *Spanish Emotion Lexicon (SEL) (Sidorov et al., 2013)*: the lexicon provides a Probability Factor of Affective use (PFA) for each of its entries, with respect to at least one of six basic emotions: joy, anger, fear, sadness, surprise and disgust.

We map emotions to a binary polarity scale, considering positive words most related to joy, and negative all the others except those related to surprise. We consider surprise an ambiguous sentiment and thus discard those words.

- *SO-CAL lexicon (Taboada et al., 2011a)* has the polarities of the words graded in a $[-5, 5]$ scale, from most negative to most positive. The less polar levels $[-3, 3]$ presented some conflicts with respect to other lexicons. Experiments were carried out in order to determine the most suitable words to be included in our lexicon.

Lexicon \Polarity	negative	positive	Total
ElhPolar	2,857	1,654	4,511
Mih (full)	476	871	1,347
SO-CAL	2,572	2,119	4,691
SEL	1,193	668	1,861 (+175 discarded)

Table 8.1: Statistics of the polarity lexicons used by our system.

8.3.3 Supervised System

We used the SMO implementation of the SVM algorithm included in the Weka (Hall et al., 2009) data mining software. All the classifiers built over the training data were evaluated by means of the 10-fold cross validation strategy. Complexity parameter was optimized ($C = 0.666667$).

Preprocessing

As mentioned in section 8.2, microblogging in general and Twitter, in particular, suffers from a high presence of spelling errors. This hampers any knowledge-based processing as well as supervised methods. Thus prior to any other process, we apply a microtext normalization step. We apply a two step normalization algorithm (Saralegi and San Vicente, 2013b). First, candidates for each unknown word are generated by means of various methods dealing with different error-sources: extension of usual abbreviations, correction of colloquial forms, correction of replication of characters, normalization of interjections, and correction of spelling errors by means of edit-distance metrics. Then, the correct candidates are selected using a language model trained on correct Spanish text corpora.

In addition, all URLs are replaced by the *"URL"* string, and text is converted to lower case (upper case information is saved for later use).

Baseline

The SVM system presented to last year's task 1 was used (Saralegi and San Vicente, 2013) as baseline. Following we give a brief overview of the features the system uses:

- *ElhPolar*: Frequency of lemmas in Elhuyar Polar polarity lexicon.
- *POS information*: the frequency of the POS tags in a message.
- *Frequency of Polarity Words (FP)*: Two features including the polarity information of the lexicon. Positivity and negativity scores of a tweet are computed based on the polarities in ElhPolar. Various phenomena, such as negation or intensity modifiers are taken into account.
- *Emoticons and Interjections*: Emoticon and interjection lists were compiled from various sources. Emoticons are grouped in 3 positive and 5 negative categories. Interjections are grouped into two classes: positive and negative interjections. Frequency of each category is included as a feature of the classifier.
- *Upper case*: Overuse of upper case (e.g., *"MIRA QUE BUENO"*) is often used to give more intensity to the tweet. The proportion of upper-cased characters in a tweet is stored as a feature.

The features described in the next sections were added on top of this initial configuration. Experiments carried out with various lexicons (section 8.3.3) influence the FP values described above.

Features / Metric	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
All features (Elh2014)	**51.54**	64.6	29.0	13.4	48.8	43.6	65.9
- Ngrams	**-0.51**	-0.9	-0.2	-0.2	-0.1	-0.9	-0.7
- Neg	**-0.13**	-0.1	-0.7	-0.1	0.3	-0.9	0.1
- Punct	-0.06	-0.2	-0.1	0.1	-0.2	0.4	0.1

Table 8.2: Ablation experiments on C_t corpus. Only the information of ElhPolar lexicon is used in these experiments. Columns 3rd to 8th show F-scores for each of the class values.

Syntax based ngrams (Ngrams)

Frequent ngram combinations can help to better identify the polarity of texts. For example, *"merecer la pena"* (to be worth), is a positive expression, but *"pena"* (pity)

is negative. Detecting such structures would be helpful for identifying prior polarities more accurately. So, we extract ngrams from the training corpus based on certain syntactic patterns. Specifically, [N+Adj] and [Verb+Noun] patterns were used to extract locutions (e.g., "perro faldero"). A minimum frequency of 3 occurrences was required for a locution to be accepted. Following this methodology, a total amount of 192 ngrams were extracted. Each of them is included as a new feature in the classifier, storing their occurrence frequency.

Experiments carried out on C_t training data-set ("- Ngrams" row in Table 8.2, indicate that those locutions are indeed helpful, specially for detecting extreme polarities (P+ and N+).

Punctuation marks (Punct)

Some authors (Proisl et al., 2013; Barbosa and Feng, 2010) suggest that punctuation marks may be good hints for detecting polarity. It is difficult to discern a specific polarity based solely on the information provided by punctuation marks, but they may be a good hint to determine intensity of the sentiment, specially when appearing at the end of a sentence. Following this intuition, we added four new features: the number of exclamation and interrogation marks in a tweet, and whether a tweet ends with and interrogation or exclamation marks.

Results on C_t show that such features do provide some improvement. Looking at the results of the training set, a single feature was included in the final configuration: whether the tweet ends with an interrogation mark or not. "- Punct" row in Table 8.2 represent the ablation study for this configuration.

Treatment of Negations (Neg)

The polarity of a word changes if it is included in a negative clause[2]. Our baseline system only takes into account negation phenomena when computing FP values. Instead, we include this information explicitly to our learning model. For each lexicon and ngram feature f, another feature NOT_f is created. This nearly duplicates the feature number used by the classifier (from 8k to 14k features).

Experiments on training data (see "- Neg" row in Table 8.2) showed that the classifier obtains a slight improvement by using those features.

[2]Syntactic information provided by FreeLing is used for detecting those cases.

Lexicons \Metric	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Elh2014 (All features)	51.54	64.6	29.0	13.4	48.8	43.6	65.9
Elh2014+SEL **(Run1)**	**51.74**	65.0	29.2	12.9	48.9	43.7	66.3
Elh2014+Mih	51.50	64.9	28.8	14.3	48.5	43.3	66.2
Elh2014+SO-CAL3	51.18	64.8	28.2	13.8	47.7	43.3	66.0
Elh2014+SEL+Mih+SO-CAL3 **(Run2)**	51.63	64.8	28.9	14.2	48.4	43.7	66.9
Elh2014+SEL+Mih+SO-CAL4 **(Run3)**	51.59	64.7	29.3	14.2	48.2	43.8	66.8

Table 8.3: Lexicon combination experiments on training data. Columns 3rd to 8th show F-scores for each of the class values.

Lexicon Combination

As we have already mentioned in section 8.3.3, FP features are the solution we have to explicitly provide the classifier with the polarity information stored in the polarity lexicons. This allows the system to take into account those polarity words not appearing in the training data. Rather than adding new influential features to the model, we expect combining lexicons will help to more accurately compute polarity score values.

Since we have combined several lexicons, conflicts arise due to words having several polarities. In order to solve those conflicts, we established a preference order. ElhPolar lexicon is first in this order, followed by SEL, SO-CAL, and Mih. We made this decision because ElhPolar is the most adapted lexicon to the corpus we are working with and it includes information extracted from the training data.

Table 8.3 presents the results of combining the various lexicons. Results are computed using all the features described in the previous sections. According to those results, neither SO-CAL nor Mih lexicons would be useful. However it is difficult to measure the real impact of such lexicons against the training data, due to the fact that most frequent polarity words in C_t are already included in the ElhPolar lexicon. That could also explain the little improvement achieved overall (0.2%). Hence, we decided to send runs for those configurations with results over the system using only ElhPolar.

Note that there are several configurations using the SO-CAL lexicon. The SO-CAL3 notation refers to using those entries in the lexicon with a polarity score > 3 or < -3. Similarly, SO-CAL4 refers to those entries with scores > 4 or < -4. Including the complete SO-CAL led to a drop in performance for us, so we

conducted experiments in order to determine if using only its most polar words could still be helpful. We only include here the configurations which achieved the best results on C_t.

8.4 Evaluation and Results

The organization provided two evaluation test-sets. On the one hand, for comparison purposes, TASS 2013's test-set C_{e2013} was used (Villena-Román et al., 2014). On the other hand, a 1,000 tweet subset was also prepared C_{e1k}, containing a more similar category distribution compared with the training corpus. Then again, it must be noted that C_{e1k} is yet more skewed towards positive polarity (50% of the whole corpus, as show in the last column of Table 8.4) and NONE tweets have been reduced considerably.

Polarity	**tweets in C_t**	**tweets in C_{e2013}**	**tweets in C_{e1k}**
P+	22.88% (1,652)	34.12% (20,745)	29.1% (291)
P	17.07% (1,232)	2.45% (1,488)	21.6% (216)
NEU	9.28% (670)	2.15% (1,305)	6.3% (63)
N	18.49% (1,335)	18.56% (11,287)	20.7% (207)
N+	11.73% (847)	7.5% (4,557)	10% (100)
NONE	20.54% (1,483)	35.22% (21,416)	12.3% (123)
Total	100% (7,219)	100% (60,798)	100% (1000)

Table 8.4: Polarity classes distribution in train and test corpora

Each participant was allowed to send up to three runs per task where 6-category classification (5 polarities + NONE) and 4-category classification (3 polarities + NONE) were considered different tasks. For the 4-category results, all tweets regarded as positive are grouped into a single category, and the same is done for negative tweets. Table 8.5 presents the results for both evaluations against the C_{e2013} corpus, using the best scored classifiers in the training process. Table 8.6 presents the results for the evaluation against the C_{e1k} data-set. In addition to the accuracy results, both tables show F-scores for each class for the 6-category classification. For the sake of readability, we will refer to our submitted systems as follows:

- **Run1**: Elh2014+SEL.

- **Run2**: Elh2014+SEL+Mih+SO-CAL3.
- **Run3**: Elh2014+SEL+Mih+SO-CAL4.

Results over the C_{e2013} data-set, show the tendency of improving the results obtained over the training set. Overall, a 1% improvement improvement is achieved over last year's system. Although the system ranked second with this corpus, it is 3% and 1% beyond the best results achieved by ELiRF-UPV team, for 6 and 4 category classifications, respectively.

Metric/ System	Acc. (4 cat.)	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Elhuyar Tass2013	68.6	60.1	72.5	22.8	14.4	54.5	46.5	66.9
Run 1	**69.9**	**61.0**	73.6	22.8	14.0	55.9	52.7	66.7
Run 2	69.7	60.6	73.1	22.7	14.6	55.8	53.0	66.1
Run 3	69.8	60.6	73.0	22.9	14.6	55.9	53.0	66.0
Best Results (ELiRF-UPV)	**70.9**	**64.3**	-	-	-	-	-	-

Table 8.5: Results obtained on the evaluation of the C_{e2013} data.

Results over the C_{e1k} data-set (see Table 8.6) are overall lower than those obtained with the C_{e2013} corpus. Accuracy is below training corpus results in all cases. However, the improvement our new features obtain over last year's system is more notable over this corpus. Also the gap between our system and the best results narrows, specially in the 6 category classification task.

It is worth mentioning that lexicon combinations' performance has a boost compared to the training data. Results on C_{e2013} (Table 8.5) behave similarly as on C_t, with run1 above the other two, although the differences are minimal, specially in the 4 category classification. In turn, Table 8.6 shows that, Mih and SO-CAL lexicons which had even a negative contribution on the training data (see Table 8.3, runs 2 and 3), provide the best results on C_{e1k} improving the results more than 1% and 2% over run 1 and baseline systems, respectively. These results remark the importance of the FP values, because many of the polarity words added by those lexicons only influence the classifier model through the FP values, because they had no occurrence in the training data.

If we take a look at the individual category results, first thing we notice is that neutral tweets are very difficult to classify. Such tweets do contain polarity words, but often they have mixed polarities. We should try to find features that better characterize such messages. The performance of negative categories drops significantly from C_{e2013} to C_{e1k}, but the results on C_{e1k} are in concordance with C_t results. NONE tweets have that same behavior on test data, but in that case, results

on C_t agree with those on C_{e2013}. In any case it is difficult for us to draw conclusions, because 43% of the tweets in C_{e1k} are annotated differently in C_{e2013}.

Metric/ System	Acc. (4 cat.)	Acc. (6 cat.)	P+	P	NEU	N	N+	NONE
Elhuyar Tass2013	61.0	44.8	66.5	24.7	14.0	40.8	39.8	45.9
Run 1	62.3	46.7	67.6	19.2	19.1	46.0	46.9	46.5
Run 2	63.2	**47.4**	67.1	23.1	19.1	46.4	45.9	48.0
Run 3	**63.5**	47.3	66.5	21.5	19.6	47.4	47.1	47.6
Best Results (ELiRF-UPV)	**65.9**	**48.0**	-	-	-	-	-	-

Table 8.6: Results obtained on the evaluation of the C_{e1k} data.

8.5 Conclusions

We have presented a SVM classifier for detecting the polarity of Spanish tweets. Our classifiers ranked second among 7 participant groups. Starting on the system developed for the TASS 2013 challenge, we have successfully incorporated a series of new features, such as syntax based ngrams or negated elements. The combination of various polarity lexicons has also contributed to the performance improvement. It must be noted that such improvement is not reflected on the experiments carried out on the training corpus. The limited size of the training data and the fact that most influential polarity words were already included in our initial lexicon, make difficult to determine to what extent the additions may help.

There is still room for improvement. We would like to further explore the use of ngram-based locutions on the one hand, for example by collecting polarity annotated locutions. On the other hand, neutral polarity is the hardest one to determine. A future line of work is to direct our efforts towards researching on how to characterize such messages.

Acknowledgments

This work has been partially founded by the Industry Department of the Basque Government under grant IE12-333 (Ber2tek project) and Spanish Government MICINN project *Skater* (Grant No. TIN2012-38584-C06-01).

CHAPTER 9

English Polarity Classification

EliXa: A modular and flexible ABSA platform

Iñaki San Vicente[1], Xabier Saralegi[1], Rodrigo Agerri[2]

[1] Elhuyar Foundation
[2] IXA NLP Group - University of the Basque Country (UPV/EHU)

This paper presents a supervised Aspect Based Sentiment Analysis (ABSA) system. Our aim is to develop a modular platform which allows to easily conduct experiments by replacing the modules or adding new features. We obtain the best result in the Opinion Target Extraction (OTE) task (slot 2) using an off-the-shelf sequence labeler. The target polarity classification (slot 3) is addressed by means of a multiclass SVM algorithm which includes lexical based features such as the polarity values obtained from domain and open polarity lexicons. The system obtains accuracies of 0.70 and 0.73 for the restaurant and laptop domain respectively, and performs second best in the out-of-domain hotel, achieving an accuracy of 0.80.

Published in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 748–752. Association for Computational Linguistics, 2015. doi: 10.18653/v1/S15-2127

9.1 Introduction

Nowadays Sentiment Analysis is proving very useful for tasks such as decision making and market analysis. The ever increasing interest is also shown in the number of related shared tasks organized: TASS (Villena-Román et al., 2012; Villena-Román et al., 2014), SemEval (Nakov et al., 2013; Pontiki et al., 2014; Rosenthal et al., 2014), or the SemSA Challenge at ESWC2014[1]. Research has also been evolving towards specific opinion elements such as entities or properties of a certain opinion target, which is also known as ABSA. The Semeval 2015 ABSA shared task aims at covering the most common problems in an ABSA task: detecting the specific topics an opinion refers to (slot1); extracting the opinion targets (slot2), combining the topic and target identification (slot1&2) and, finally, computing the polarity of the identified word/targets (slot3). Participants were allowed to send one constrained (no external resources allowed) and one unconstrained run for each subtask. We participated in the slot2 and slot3 subtasks.

Our main is to develop an ABSA system to be used in the future for further experimentation. Thus, rather than focusing on tuning the different modules our goal is to develop a platform to facilitate future experimentation. The EliXa system consists of three independent supervised modules based on the IXA pipes tools (Agerri et al., 2014) and Weka (Hall et al., 2009). Next section describes the external resources used in the unconstrained systems. Sections 9.3 and 9.4 describe the systems developed for each subtask and briefly discuss the obtained results.

9.2 External Resources

Several polarity Lexicons and various corpora were used for the unconstrained versions of our systems. To facilitate reproducibility of results, every resource listed here is publicly available.

9.2.1 Corpora

For the restaurant domain we used the Yelp Dataset Challenge dataset[2]. Following (Kiritchenko et al., 2014), we manually filtered out categories not corresponding to

[1] http://challenges.2014.eswc-conferences.org/index.php/SemSA
[2] http://www.yelp.com/dataset_challenge

food related businesses (173 out of 720 were finally selected). A total of 997,721 reviews (117.1M tokens) comprise what we henceforth call the *Yelp food corpus* (C_{Yelp}).

For the laptop domain we leveraged a corpus composed of Amazon reviews of electronic devices (Jo and Oh, 2011). Although only 17,53% of the reviews belong to laptop products, early experiments showed the advantage of using the full corpus for both slot 2 and slot 3 subtasks. The *Amazon electronics corpus* (C_{Amazon}) consists of 24,259 reviews (4.4M tokens). Finally, the English Wikipedia was also used to induce word clusters using word2vec (Mikolov et al., 2013).

9.2.2 Polarity Lexicons

We generated two types of polarity lexicons to represent polarity in the slot3 subtasks: general purpose and domain specific polarity lexicons.

A general purpose polarity lexicon L_{gen} was built by combining four well known polarity lexicons: SentiWordnet SWN (Baccianella et al., 2010), General Inquirer GI (Stone et al., 1966), Opinion Finder OF (Wilson et al., 2005) and Liu's sentiment lexicon Liu (Hu and Liu, 2004b). When a lemma occurs in several lexicons, its polarity is solved according to the following priority order: $Liu > OF > GI > SWN$. The order was set based on the results of (San Vicente et al., 2014). All polarity weights were normalized to a $[-1, 1]$ interval. Polarity categories were mapped to weights for GI ($neg_+ \rightarrow -0.8$; $neg \rightarrow$-0.6; $neg_- \rightarrow$-0.2; $pos_- \rightarrow 0.2$; $pos \rightarrow 0.6$; $pos_+ \rightarrow 0.8$), Liu and OF ($neg \rightarrow$-0.7; $pos \rightarrow 0.7$ for both). In addition, a restricted lexicon L_{genres} including only the strongest polarity words was derived from L_{gen} by applying a threshold of ± 0.6.

Domain	Polarity Lexicon	Total
General	L_{gen}	42,218
General	L_{genres}	12,398
Electronic devices	L_{Amazon}	4,511
Food	L_{Yelp}	4,691

Table 9.1: Statistics of the polarity lexicons.

Domain specific polarity lexicons L_{Yelp} and L_{Amazon} were automatically extracted from C_{Yelp} and C_{Amazon} reviews corpora. Reviews are rated in a [1..5] interval,

being 1 the most negative and 5 the most positive. Using the Log-likelihood ratio (LLR) (Dunning, 1993) we obtained the ranking of the words which occur more with negative and positive reviews respectively. We considered reviews with 1 and 2 rating as negative and those with 4 and 5 ratings as positive. LLR scores were normalized to a $[-1, 1]$ interval and included in L_{Yelp} and L_{Amazon} lexicons as polarity weights.

9.3 Slot2 Subtask: Opinion Target Extraction

The Opinion Target Extraction task (OTE) is addressed as a sequence labeling problem. We use the *ixa-pipe-nerc* Named Entity Recognition system[3] (Agerri et al., 2014) off-the-shelf to train our OTE models; the system learns supervised models via the Perceptron algorithm as described by Collins (2002). *ixa-pipe-nerc* uses the Apache OpenNLP project implementation of the Perceptron algorithm[4] customized with its own features. Specifically, *ixa-pipe-nerc* implements basic non-linguistic local features and on top of those a combination of word class representation features partially inspired by Turian et al. (2010). The word representation features use large amounts of unlabeled data. The result is a quite simple but competitive system which obtains the best constrained and unconstrained results and the first and third best overall results.

The local features implemented are: current token and token shape (digits, lowercase, punctuation, etc.) in a 2 range window, previous prediction, beginning of sentence, 4 characters in prefix and suffix, bigrams and trigrams (token and shape). On top of them we induce three types of word representations:

- Brown clusters (Brown et al., 1992), taking the 4th, 8th, 12th and 20th node in the path. We induced 1000 clusters on the Yelp reviews dataset described in section 9.2.1 using the tool implemented by Liang[5].

- Clark clusters (Clark, 2003), using the standard configuration to induce 200 clusters on the Yelp reviews dataset and 100 clusters on the food portion of the Yelp reviews dataset.

[3]https://github.com/ixa-ehu/ixa-pipe-nerc
[4]http://opennlp.apache.org/
[5]https://github.com/percyliang/brown-cluster

- Word2vec clusters (Mikolov et al., 2013), based on K-means applied over the extracted word vectors using the skip-gram algorithm[6]; 400 clusters were induced using the Wikipedia.

The implementation of the clustering features looks for the cluster class of the incoming token in one or more of the clustering lexicons induced following the three methods listed above. If found, then we add the class as a feature. The Brown clusters only apply to the token related features, which are duplicated. We chose the best combination of features using 5-fold cross validation, obtaining 73.03 F1 score with local features (e.g. constrained mode) and 77.12 adding the word clustering features, namely, in unconstrained mode. These two configurations were used to process the test set in this task. Table 9.2 lists the official results for the first 4 systems in the task.

System (type)	Precision	Recall	F1 score
Baseline	55.42	43.4	48.68
EliXa (u)	68.93	71.22	**70.05**
NLANGP (u)	70.53	64.02	67.12
EliXa (c)	67.23	66.61	66.91
IHS-RD-Belarus (c)	67.58	59.23	63.13

Table 9.2: Results obtained on the slot2 evaluation on restaurant data.

The results show that leveraging unlabeled text is helpful in the OTE task, obtaining an increase of 7 points in recall. It is also worth mentioning that our constrained system (using non-linguistic local features) performs very closely to the second best overall system by the NLANGP team (unconstrained). Finally, we would like to point out to the overall low results in this task (for example, compared to the 2014 edition), due to the very small and difficult training set (e.g., containing many short samples such as "Tasty Dog!") which made it extremely hard to learn good models for this task. The OTE models will be made freely available in the *ixa-pipe-nerc* website in time for SemEval 2015.

[6]https://code.google.com/p/word2vec/

9.4 Slot3 Subtask: Sentiment Polarity

The EliXa system implements a single multiclass SVM classifier. We use the SMO implementation provided by the Weka library (Hall et al., 2009). All the classifiers built over the training data were evaluated via 10-fold cross validation. The complexity parameter was optimized as ($C = 1.0$). Many configurations were tested in this experiments, but in the following we only will describe the final setting.

Classifier	Acc Rest
Baseline (organizers)	78.8
Baseline	
1lgram	80.11
2lgram	79.3
$1lgram + E\&A$	79.8
$1lgram(w5)$	80.41
$1lgram + PoS$	80.59 (c)
Lexicons	
$1lgram + L_{gen}$	80.6
$1lgram + L_{genres}$	81
$1lgram + L_{Yelp}$	80.9
Combinations	
$1lgram(w5) + w2v(C_{Yelp}) + L_{genres} + L_{Yelp} + PoS$	82.34 (u)

Table 9.3: Slot3 ablation experiments for restaurants. (c) and (u) refer to constrained and unconstrained tracks.

9.4.1 Baseline

The very first features we introduced in our classifier were token ngrams. Initial experiments showed that lemma ngrams (lgrams) performed better than raw form ngrams. One feature per lgram is added to the vector representation, and lemma frequency is stored. With respect to the ngram size used, we tested up to 4-gram features and improvement was achieved in laptop domain but only when not combined with other features.

Classifier	Acc Lapt
Baseline (organizers)	78.3
Baseline	
1lgram	79.33
2lgram	79.7
$1lgram + clusters(w2v)$	79.23
$1lgram + E\&A$	79.23
$1lgram + PoS$	78.88
Lexicons	
$1lgram + L_{gen}$	79.2
$1lgram + L_{genres}$	79
$1lgram + L_{Amazon}$	79.7
Combinations	
$1lgram + PoS + E\&A$	79.99 (c)
$2lgram + PoS + E\&A$	78.27
$1lgram + L_{genres} + L_{Amazon} + PoS + E\&A$	80.85 (u)

Table 9.4: Slot3 ablation experiments for laptops. (c) and (u) refer to constrained and unconstrained tracks.

9.4.2 POS

POS tag and lemma information, obtained using the IXA pipes tools (Agerri et al., 2014), were also included as features. One feature per POS tag was added again storing the number of occurrences of a tag in the sentence. These features slightly improve over the baseline only in the restaurant domain.

9.4.3 Window

Given that a sentence may contain multiple opinions, we define a window span around a given opinion target (5 words before and 5 words after). When the target of an opinion is null the whole sentence is taken as span. Only the restaurant and hotel domains contained gold target annotations so we did not use this feature in the laptop domain.

9.4.4 Polarity Lexicons

The positive and negative scores we extracted as features from both general purpose and domain specific lexicons. Both scores are calculated as the sum of every positive/negative score in the corresponding lexicon divided by the number of words in the sentence. Features obtained from the general lexicons provide a slight improvement. L_{genres} is better for restaurant domain, while L_{gen} is better for laptops. Domain specific lexicons L_{Amazon} and L_{Yelp} also help as shown by tables 9.3 and 9.4.

9.4.5 Word Clusters

Word2vec clustering features combine best with the rest as shown by table 9.3. These features only were useful for the restaurant domain, perhaps due to the small size of the laptops domain data.

9.4.6 Feature combinations

Every feature, when used in isolation, only marginally improves the baseline. Some of them, such as the E&A features (using the gold information from the slot1 subtask) for the laptop domain, only help when combined with others. Best performance is achieved when several features are combined. As shown by tables 9.3 and 9.4, improvement over the baseline ranges between 2,8% and 1,9% in the laptop and restaurant domains respectively.

9.4.7 Results

Table 9.5 shows the result achieved by our sentiment polarity classifier. Although for both restaurant and laptops domains we obtain results over the baseline both performance are modest.

In contrast, for the out of domain track, which was evaluated on hotel reviews our system obtains the third highest score. Because of the similarity of the domains, we straightforwardly applied our restaurant domain models. The good results of the constrained system could mean that the feature combination used may be robust across domains. With respect to the unconstrained system, we suspect that such a good performance is achieved due to the fact that word cluster information was very adequate for the hotel domain, because C_{yelp} contains a 10.55% of hotel reviews.

System	Rest.	Lapt.	Hotel
Baseline	63.55	69.97	71.68 (majority)
Sentiue	**78.70 (1)**	**79.35 (1)**	71.68 (4)
lsislif	75.50 (3)	77.87 (3)	**85.84 (1)**
EliXa (u)	70.06(10)	72.92 (7)	79.65 (3)
EliXa (c)	67.34 (14)	71.55 (9)	74.93 (5)

Table 9.5: Results obtained on the slot3 evaluation on restaurant data; ranking in brackets.

9.5 Conclusions

We have presented a modular and supervised ABSA platform developed to facilitate future experimentation in the field. We submitted runs corresponding to the slot2 and slot3 subtasks, obtaining competitive results. In particular, we obtained the best results in slot2 (OTE) and for slot3 we obtain 3rd best result in the out-of-domain track, which is nice for a supervised system. Finally, a system for topic detection (slot1) is currently under development.

Acknowledgments

This work has been supported by the following projects: ADi project (Etortek grant No. IE-14-382), NewsReader (FP7-ICT 2011-8-316404), SKaTer (TIN2012-38584-C06-02) and Tacardi (TIN2012-38523-C02-01).

PART IV

REAL WORLD APPLICATION

IV Real World application

Up to this point this manuscript has presented the experimentation path we followed in order to develop the different modules required to develop a multilingual Sentiment Analysis system. However, we still have to assemble the complete jigsaw puzzle with the components developed so far.

One of our main goals was to generate the first supervised sentiment analysis system for Basque. We have lexicons for Basque (see Part I), and a trainable system, but we still lack annotated data at polarity level. Furthermore, we also lack tweet normalization resources for Basque.

The final part of this thesis describes how to combine the previously acquired knowledge into a real world application: Talaia. Talaia is a platform that allows automatic analysis of the impact in social media and digital press of topics or domains specified by the user. This application is the culmination of all the previously presented work. Talaia was first developed in the framework of Behagunea[7], a project to monitor in real time the activity around the European cultural capital of Donostia 2016. In fact, the creation of the multilingual resources for all four languages mentioned across this thesis were developed in the framework of Behagunea. Thus, polarity lexicons, polarity annotated tweet datasets, polarity classification models and tweet normalization resources were developed in four languages: Basque, English, French and Spanish.

A single paper is presented for this part:

- Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. Real time monitoring of social media and digital press. submitted to Engineering Applications of Artificial Intelligence journal, Elsevier. ISSN: 0952-1976. Preprint available at `https://arxiv.org/abs/1810.00647`, 2019

[7] `https://sustatu.eus/aitzol_astigarraga/1465395090`

Contribution to the paper: Main author of the paper. Iñaki San Vicente is the main developer of the application presented, including the crawler, the polarity classifier and the data visualization interface. He is also responsible for compiling the resources presented in the paper, including the polarity classification models. Several people contributed to the creation of datasets. Xabier Saralegi contributed designing the architecture of the platform and the crawling modules. Rodrigo Agerri contributed to the evaluation of the classifiers and the writing of the paper.

CHAPTER 10

Social Media Sentiment Monitor

Real time Monitoring of Social Media and Digital Press

Iñaki San Vicente[1], Xabier Saralegi[1], Rodrigo Agerri[2]

[1] Elhuyar Foundation
[2] IXA NLP Group - University of the Basque Country (UPV/EHU)

Talaia is a platform for monitoring social media and digital press. A configurable crawler gathers content with respect to user defined domains or topics. Crawled data is processed by means of the EliXa Sentiment Analysis system. A Django powered interface provides data visualization for a user-based analysis of the data. This paper presents the architecture of the system and describes in detail its different components. To prove the validity of the approach, two real use cases are accounted for: one in the cultural domain and one in the political domain. Evaluation for the sentiment analysis task in both scenarios is also provided, showing the capacity for domain adaptation.

Submitted to *Engineering Applications of Artificial Intelligence*. Elsevier. ISSN: 0952-1976. Preprint digital version available at `https://arxiv.org/abs/1810.00647`

10.1 Introduction

The Internet is a very rich source of user-generated information. As knowledge management technologies have evolved, many organizations have turned their eyes to such information, as a way of obtaining global feedback on their activities (Chen et al., 2012). Some studies (O'Connor et al., 2010; Ceron et al., 2015) have pointed out that such systems could perform as well as traditional polling systems, but at a much lower cost.

Talaia is a platform for monitoring the impact of topics specified by the user in social media and digital press. The process starts when the user configures the system to find information related to a domain or topic. Talaia provides real time information on the topic and graphic visualizations to help users interpreting the data. Such technology has various applications areas, such as:

- Monitoring events: Following public events in real time harvesting people's opinions and media news (Sutton, 2009; Yu and Wang, 2015).
- Analyze citizen or electors voice: Tracking the opinions citizens convey with respect to public services or trends during electoral campaigns (Ceron et al., 2015).
- Marketing and brand management: Measuring the impact of marketing campaigns in a digital environment (Ahmed et al., 2018).
- Business Intelligence: Fast and efficient visualization of the information extracted from social media offers companies the possibility to analyze opinions about their products or services (He et al., 2013; Mostafa, 2013).
- Security: Detection of social conflicts, crimes, and cyberbullying (Xu et al., 2012; Dadvar et al., 2013).

Talaia consists of three main modules: (i) a crawler collecting the data; (ii) a data analysis module; and (iii) a Graphical User Interface (GUI) providing interpretation of the data analyzed. Figure 10.1 describes the architecture. Its main features are the following:

- Monitoring and automatic analysis: Definition of the domain/topic by means of term taxonomies. Continuous monitoring of various mention sources including social media and digital press.

- Multilingual extraction of mentions and opinions relevant to the topics monitored, by means of Natural Language Processing (NLP) techniques.
- Result exploration: Intuitive GUI to visualize and analyze the results. Advanced statistics and filters, such as per language results, impact of the topics or author statistics.
- Control of the monitoring process through the user interface: update search terms or review and correct gathered mentions.

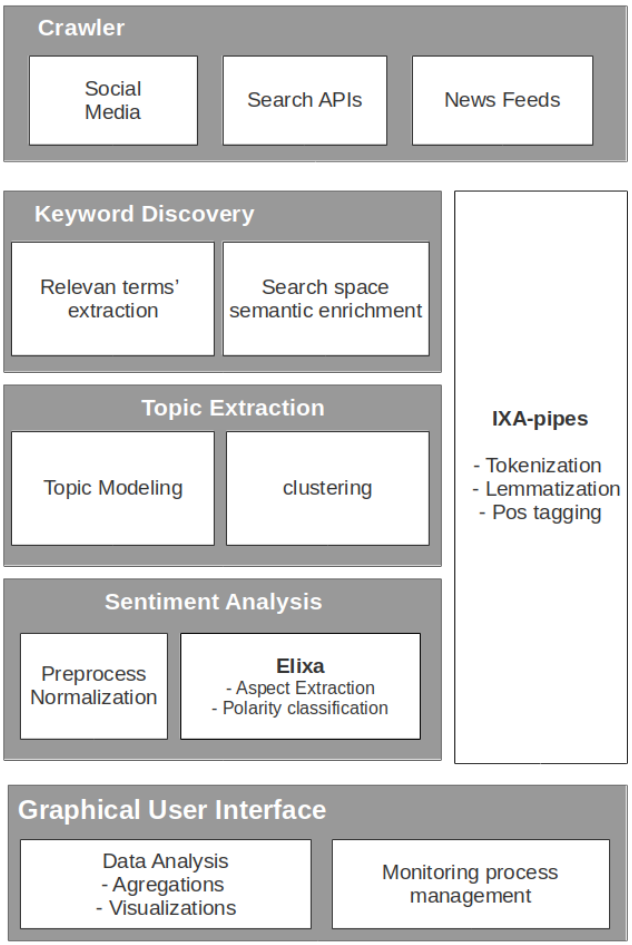


Figure 10.1: Diagram showing Talaia's components and architecture.

This paper focuses on those processes monitoring user satisfaction with respect to a topic, and that is why we pay special attention to the Sentiment Analysis (SA) module. Nevertheless, Talaia is capable of performing further data analysis tasks involving user profiling, in order to get the most out of the data. Specifically, geolocalization, user community identification and gender detection have been implemented. Section 10.4.4 provides more details.

The rest of the paper is organized as follows. Section 10.2 discusses previous work, focusing on social media and SA. Both the academic and industrial points of view are taken into account. The third section describes in detail the modules composing Talaia. Section 10.6 presents two success cases where the platform has been used for monitoring different events. Section 10.7 provides evaluation and results on the SA task for both scenarios. The last section draws some conclusions and future directions.

10.2 Background

10.2.1 Social Media Analysis

Social media are becoming the primary environment for producing, spreading and consuming information. Enormous quantities of user generated content are produced constantly. Even traditional media spread their news and get a large amount of traffic trough social media. Monitoring events or topics in such an environment is however a challenging task. That is where data mining and Natural Language Processing (NLP) become essential. We have to be able to collect large scale data, but also to extract the relevant information. Tracking a topic over an extended time period means that the information flow grows and fades over time. Also a topic may evolve in terms of the vocabulary used, and thus "topic detection and tracking" (TDT) (Allan et al., 1998) techniques become relevant to maintain a successful monitoring.

Several systems have been proposed in the literature to explore events. Trend Miner (Preoţiuc-Pietro and Cohn, 2013) extracts multilingual terms from social media, groups and visualizes them in temporal series. Social Sensor (Aiello et al., 2013) and Twitcident (Abel et al., 2012) may be the most similar systems to ours. The first one focuses on tracking topic or events predefined by the user. The second makes user defined searches related to crisis management. LRA[1] aims to discovering and tracking crisis situations based on crowdsourced information. ReDites (Osborne et al., 2014) detects an tracks topics in a fully automated way.

[1] `https://www.lracrisistracker.com`

Detecting terms that represent a domain or topic semantically has been traditionally addressed by statistical models such as Latent Dirichlet Association (LDA) (Blei et al., 2003). Classical LDA models are applied over static document collections. In order to extract terms from dynamic collections, the most common approach is to follow a two step strategy (Shamma et al., 2011) consisting of detecting emerging terms and grouping them in clusters thereby defining a domain.

Nguyen et al. (2016) predict emerging terms by means of word co-occurrence distributional models, comparing the terms in an specific time window against the whole collection. Abilhoa and De Castro (2014) use a graph-based representation of the document collection. Aiello et al. (2013) propose $df\text{-}idf_t$ (Document Frequency - Inverse Document Frequency), a variation of $tf\text{-}idf$ that includes the temporal factor. Kim et al. (2016) combine neural networks and sequence labelling in order to extract relevant terms from conversations. Miao et al. (2017) propose to reduce the cost of predicting emerging topics by finding a small group of representative users and predict the emerging topics from their social media activity.

There is also the problem of the scope of the event or topic to be tracked. An event might be tracked at global level (e.g. Football World Cup), but most events are local or regional at most. Two issues arise at this point. Firstly, how to restrict the data gathered to a specific region, and, secondly, how to cope with multilingual data. Some authors tackle the problem by automatically geolocating tweets while others try focus on user locations. See Jurgens et al. (2015); Zubiaga et al. (2017) for reviews of previous approaches. Our approach is to geolocate users rather than tweets, in order to construct a census of Twitter users in a region. We developed a SVM classifier similar to Zubiaga et al. (2017) using follower and friend information as features.

10.2.2 Sentiment Analysis

In the last years microblogging sites such as Twitter have attracted the attention of many researchers with diverse objectives such as stock market prediction (Bollen et al., 2010; Oliveira et al., 2017), polling estimation (O'Connor et al., 2010; Ceron et al., 2015) or analysis of crisis situations (Pope and Griffith, 2016; Shaikh et al., 2017; Öztürk and Ayvaz, 2018). The growing number of SA related shared tasks (e.g., SemEval Aspect based SA and Twitter SA shared tasks) or the commercial platforms for reputation management (see section 10.2.3) are proof of the interest from both academic and market worlds.

The particularities of its language make it hard to analyze tweets. User mentions, hashtags, the growing presence of emojis, ungrammatical sentences, vocabulary variations and other phenomena pose a great challenge for traditional NLP tools (Foster et al., 2011; Liu et al., 2011). Brody and Diakopoulos (2011) deal with the word lengthening phenomenon, which is especially important for sentiment analysis because it usually expresses emphasis of the message. Hashtag decomposition (e.g.,*#GameOfThrones = 'Game Of Thrones'*) (Brun and Roux, 2014; Belainine et al., 2016) or matching Out Of Vocabulary (OOV) forms and acronyms to their standard vocabulary forms (e.g., *'imo = in my opinion'*) (Han and Baldwin, 2011; Liu et al., 2012; Alegria et al., 2014) are other addressed issues. International benchmarking initiatives such as the TweetNorm shared task(Alegria et al., 2015) or the WNUT[2] workshop series are proof of the interest to solve this task.

Once texts are normalized, sentiment analysis can be performed. Several ruled-based systems to polarity classification have been proposed (Hu and Liu, 2004b; Thelwall, 2017; Taboada et al., 2011a). Nevertheless, we will focus on Machine Learning (ML) based approaches which are the most widespread. Support Vector Machines (SVM) and Logistic Regression algorithms have been the very popular for polarity classification as various international shared tasks (Román et al., 2015; Pontiki et al., 2014; Rosenthal et al., 2014) show. Typical features of those systems include sentiment word/lemma ngram features, POS tags (Barbosa and Feng, 2010), Sentiment Lexicons (Kouloumpis et al., 2011), emoticons (O'Connor et al., 2010), discourse information (Somasundaran et al., 2009) or, more recently, word embeddings (Mikolov et al., 2013) and clusters (San Vicente et al., 2015).

From 2015 onwards, the academic world has shifted to Deep Learning (DL) approaches, as Nakov et al. (2016) confirm. Long-Short Term Memory (LSTM) Recurrent Neural Networks (RNN)(Dai and Le, 2015; Johnson and Zhang, 2016) and Convolutional Neural Networks (CNN) are the preferred choices. Severyn and Moschitti (2015) use a single layer CNN, first to construct word embeddings and then to train the classifier. Deriu et al. (2017) propose a two phase training method: first they train a neural network with large amounts of weakly supervised data collected from Twitter. The network is initialized with word embeddings learned by means of word2vec (Mikolov et al., 2013) from very large corpora collected from twitter. Second, the weights learned in the first step are transferred to a second neural network trained over the actual annotated data, to learn the final classifier. A two convolutional layer CNN is used for both training phases. A very similar approach is followed by Cliche (2017) which achieves top results in SemEval (Rosenthal et al.,

[2] `http://noisy-text.github.io/2018/`

2017b). Howard and Ruder (2018) follow a similar three step approach with a more complex network topology obtaining state of the art results for various tasks, including sentiment analysis.

A common problem of supervised approaches, specially of DL, is the need of large amounts of labeled data for training. The common practice in the literature is to gather weakly supervised datasets following the emoticon heuristic[3] (Go et al., 2009). This is feasible for major languages, but it is a very difficult (if possible at all) and time costly task for non major languages such as Basque.

Our system is closest to Barbosa and Feng (2010) and Kouloumpis et al. (2011) because it combines polarity lexicons with machine learning for labelling sentiment of tweets. This strategy has proven to be a successful approach in previous shared tasks (Saralegi and San Vicente, 2012; Mohammad et al., 2013).

10.2.3 Industrial Solutions

We can find various commercial solutions in the market. We are particularly interested in systems that provide an integral solution of the monitoring process, leaving out tools that only approach specific phases of the surveillance process, or solutions that offer bare NLP processing chains which require further development to achieve a working social media monitor. Table 10.6 in Annex I offers a detailed comparative of the tools analyzed. We focus our analysis on the sources where information is gathered on, their tracking capabilities, the processing of multilingual information, and the data visualization.

Iconoce[4] is a system oriented to reputation management, offering various features such as measuring impact of campaigns, or reputation monitoring. Although it also can monitor social media (Twitter and Facebook) its strength lies on the analysis of digital press. Multilingual information can be gathered but no linguistic processing is performed (lemmatization or crosslingual searches). It has 3 separated search engines for authors, mentions and comments. A customizable dashboard offers various visualizations and data aggregations (e.g., salient term and topics, influencer, sentiment or trends). Periodical reports and alerts in the face of tendency changes are provided. As a distinctive feature, it offers a personalized press archive based on the customer configuration.

[3]Collect tweets containing the ":)" emoticon and regard them as positive, and likewise for the ":(" emoticon.

[4]`http://info.iconoce.com/`

In a similar way, **INNGUMA**[5] is a tool providing business intelligence services. They put their main effort in the crawling step. Rather than offering to the user results over analyzed data, the tool is designed for a group of customers to analyze the data collaboratively. Customers are provided with a search engine (more or less powerful depending on the pricing plan), and interface where they can store and share their findings.

Lexalitycs[6] and **Meaning Cloud**[7] are text analytics enterprises. Their strength is the data analysis part rather than the monitoring of many sources. Both systems are built upon robust NLP chains. Document classification, entity extraction and aspect based sentiment analysis are performed. Sentiment Analysis is approached by means of rule-based systems based on lexicons and deep linguistic analysis, offering the possibility of custom domain adaptations. Both Lexalitics and Meaning Cloud lack a result visualization interface, limiting their outputs to Excel plugins, leaving the full analysis of the data into the user's hands.

Websays[8] monitors a wide range of sources including news, Blogs/RSS, Forums, Facebook, Twitter, LinkedIn, Instagram, Foursquare, Pinterest, Youtube, Vimeo, Reviews (Tripadvisor, Booking,...). The user is able to configure the crawling using keywords. Negative words are also allowed in order to effectively restrict the search to the desired domain. The system is able to process data in several languages, but they report to be most effective with European languages (Spanish, English, French, Italian, and Catalan). SA is performed by combining ML algorithms and human validation, so the statistical models may learn from corrected data. The user may navigate through results using a dashboard that offers multiple filtering options. Graphs, salient terms, trending topics, influencers, sentiment and trends are provided, as well as periodical alerts and reports. The interface offers the possibility to manually edit and correct the results.

Following the same concept of Websays, **Keyhole**[9] is a monitoring and analytics tool that provides trends, insights, and analysis (including sentiment) of hashtags, keywords, or accounts on Twitter and Instagram. It reports supporting data processing in a number of languages, but no details are given on the technology. User can also track web mentions, but two separate monitoring processes must be setup.

[5] `https://www.innguma.com`
[6] `https://www.lexalytics.com`
[7] `https://www.meaningcloud.com/`
[8] `https://websays.com/`
[9] `https://keyhole.co/`

Lynguo[10] is also in the same group of Websays and Keyhole. It claims to provide support in 24 languages, although it reports full processing chain for Spanish and English[11]. NLP is done by means of "a range of linguistic tools to cover and combine in real time the different lexical, morphological and semantic processing layers, with machine learning and deep learning models, and software architectures"[12]. SA includes lexicons, customizable by the user. Monitoring is configured specifying keywords and users, allowing for negative ones as well. Lynguo is also able to geolocate comments.

Ubermetrics[13] is one of the few platforms that monitors multimedia sources including Youtube and Vimeo, but also TV and Radio sources. According to their reports, it processes data in 40 languages. Its visualization dashboard offers customizable graphs based on multiple search criteria. Ubermetrics main objective is analyzing virality (impact) of the mentions and author profiling.

Snaptrends[14] monitors social media (Twitter, Facebook, Instagram, and Pinterest). Multilingual data is handled by means of MT (80 languages to English). It uses a proprietary NLP chain for processing English data, including sentiment analysis and relevant term extraction. The main feature for filtering large volumes of information is a geolocation-based search engine, combined with keyword based searches and other filters such as data sources. With respect to visualization, it has various data aggregations, such as influencer rankings or sentiment evolution across time by geographical area. Snaptrend makes an special effort in visualizing specific data, generating mention mosaics and timelines in real time.

Talaia shares features with many of the aforementioned commercial solutions, yet it also possess its own characteristics. With a more robust text analysis than Iconoce and INNGUMA, and a more advanced interpretation of the data than MeaningCloud and Lexalitics, Talaia is closer to tools such as Websays and Lynguo. Having keywords organized in a taxonomy allows us to provide deeper data analysis and aggregations. Moreover, Talaia is built using open source software with a strong academical background and tested against well known benchmarks. Talaia's performance is thus, verifiable.

[10] http://lynguo.iic.uam.es/

[11] http://www.iic.uam.es/en/big-data-services/digital-environment/lynguo-en/

[12] http://www.iic.uam.es/en/big-data-services/customer-intelligence-environment/natural-language-processing/

[13] https://www.ubermetrics-technologies.com/

[14] http://snaptrends.com/

10.3 Data Collection

The first step of a monitoring system such as Talaia is the collection of information. The Multi Source Monitor (MSM) system[15] is currently able to monitor Twitter, syndication feeds and also multimedia sources such as television or radio programs. Support for other social media such as Youtube, Facebook, etc. is under development.

MSM is a keyword based crawler, which works on a set of keywords defined by the user. Rather than a list of unconnected terms, Talaia is designed to work over a hierarchy, which allows a better organization of the data for the analysis step. In this way, keywords are defined as belonging to a specific category in the taxonomy. One handicap of crawling using a keyword-based strategy is that it is often difficult to define unambiguous terms that do not capture noisy messages. In order to minimize this situation, MSM implements a number of features:

1. **Regular expressions** are used to define keywords. This allows to differentiate between common words and proper names, or full words and affixes (e.g., *podemos* 'we can' vs. *Podemos* political party). These phenomena are specially frequent in social media, where language rules are often ignored.

2. **Language specific keywords**. A word that is a very good keyword in a language can be a source of noise in another, e.g. *mendia*, 'mountain' in Spanish, is unambiguously referring to 'Idoia Mendia', a Basque politician, in our context, while in Basque it is clearly ambiguous.

3. **Anchor terms** usually define the general topic (e.g. election campaign) to monitor. If the user specifies that a keyword requires an anchor, then in order to accept a message containing that keyword the message must also contain at least one anchor term. Anchor terms may be keywords or not.

4. **Long paragraphs are split** before looking for keywords in the case of messages coming from news sites. First, it looks if any keyword appears in a candidate article. If so, it looks for keywords sentence by sentence, and those sentences are considered as the message unit.

Language Identification (LID)

LID is indispensable in order to apply the corresponding NLP analysis. LID is integrated into the crawling process as part of the MSM system. There are two

[15] `http://github.com/Elhuyar/MSM`

main reasons for that. First, it allows us implement the aforementioned "language specific keyword" feature. Second, having the language identified in the first place gives us flexibility for applying the subsequent NLP tools. At the moment language identification is implemented using the library Optimaize[16], combined with source specific optimizations (social media vs. feeds).

10.4 Data Analysis

The data analysis is mainly performed by EliXa (San Vicente et al., 2015) which integrates the following processes, each of them further detailed in the next sections.

EliXa[17] is a supervised Sentiment Analysis system. It was developed as a modular platform which allows to easily conduct experiments by replacing the modules or adding new features. It was first tested in the ABSA 2015 shared task at SemEval workshop (Pontiki et al., 2015). EliXa currently offers resources and models for 4 languages: Basque, Spanish, English and French. Its implementation is easily adaptable to other languages, requiring a polarity lexicon and/or a training dataset for each new language.

10.4.1 Normalization

To address the particular characteristics of tweets, EliXa integrates a microtext normalization module which is applied to social media messages, based on Saralegi and San Vicente (2013b). The normalizer is based on heuristic rules, such as standardizing URLs, normalizing character repetitions or dividing long words (e.g. #AVeryLongDay → 'a very long day'). Also Out Of Vocabulary (OOV) term normalization is addressed by means of language specific frequency lists based on Twitter corpora.

Furthermore, EliXa's normalization component also includes various specific functionalities related to SA:

- Emoticons are normalized into a 7 sentiment scale: *smiley, crying, shock, mute, angry, kiss, sadness.*

[16]https://github.com/optimaize/language-detector
[17]https://github.com/Elhuyar/Elixa

- Expressions that are meaningful for detecting SA such as interjections and onomatopoeia are marked.

Those normalized terms must be included in the polarity lexicons in order to have a greater impact in the sentiment analysis classification. Table 10.1 presents the resources provided for normalization according to their use. Word form dictionaries are composed of word forms extracted from corpora. When applying microtext normalization, candidates are compared to forms in the dict in order to discard noisy candidates. For example, if we were to normalize "happppy", we would know the that the correct normalization is "happy" by looking at theses dictionaries.

OOV dictionaries are composed of "OOV - standard form" pairs. These resources are valuable to normalize slang and commonly used abbreviations. In order to produce such dictionaries word form frequency lists were generated from Twitter corpora, and after pruning standard dictionary forms, the most frequent n forms were manually reviewed and manually translated[18]. When available, dictionaries were completed using precompiled lists existing in the Web.

Emoticon lexicon is a dictionary of regular expression matching a number of emoticons to their corresponding sentiment in the aforementioned scale.

Lastly, stopword lemma lists are used to discard most frequent lemmas when extracting n-gram features from texts. We adapted this lists to SA requirements by removing some lemmas, because of their relevance to polarity classification (e.g., no, good, ...).

10.4.2 NLP pre-processing

EliXa currently performs tokenization, lemmatization and POS tagging prior to sentiment analysis classification. No entity recognition is applied; entities are matched only if they are defined as keywords. Although EliXa is able to work with corpora preprocessed with other taggers, its default NLP processing is made by means of IXA pipes (Agerri et al., 2014) which is integrated as a library.

10.4.3 Sentiment Analysis

EliXa's core feature is its polarity classifier, which implements a multiclass Support Vector Machine (SVM) algorithm (Hall et al., 2009) combining the information

[18] n varies depending on the size of the input corpus. We reviewed up to 1,500 candidates.

Resource	Use	Language			
		eu	es	en	fr
Word form dictionaries	text normalization (e.g. 4ever→forever)	122,085	556,501	67,811	453,037
OOV dictionaries	text normalization (e.g. 4ever→forever)	63	7,823	223	279
Emoticon lexicon	Polarity tagging	60 (regexes matching emoji groups)			
Stopword lemma lists	Polarity tagging feature extraction	56	46	75	100

Table 10.1: Resources for text normalization included in EliXa.

extracted from polarity lexicons with linguistic features obtained from the NLP pre-processing step. Main features include polarity values from general and domain specific polarity lexicons, lemma and POS tag ngrams and positivity and negativity counts based on polarity lexicons. Features representing other linguistic phenomena such as treatment of negation, locutions or punctuation marks are also included. Finally, there are some social media specific features, such as the proportion of capitalized symbols (which often is used to increase the intensity of the message) or emoticon information.

EliXa currently provides ready to use polarity classification models, although one of its strengths is that new models can easily be trained if training data is available for a new domain.

10.4.4 User profiling

Talaia is also capable of providing deeper analysis of the data, by means of user profiling. Specifically, geolocation, gender detection and user community identification are implemented.

Opinions gathered are geolocated. Geolocation may be done using two different approaches: (i) building a census of twitter users on a region or (ii) trying to geolocate the origin of the users detected. Approach (i) is most suitable when monitoring is done for a specific region, and high precision is required from geolocation. Details on this approach are given in section 10.6.2.

Approach (ii) allows Talaia to analyze the differences in opinions with respect to a topic that may arise between regions or countries. Geolocation is done by exploiting social media information from both messages and authors. If a message is geolocated, its information is used straightforwardly. Otherwise, user profile information is analyzed. The task is challenging, because users do not provide such information always (40%-45% of the users in our datasets have location information), or they define fictitious locations (e.g., 'Middle earth', 'In a galaxy far, far away...'; 14%). Several geocoding APIs[19] are queried, and results are then weighted, because APIs show divergent results when feeding fictitious locations. The weighted system obtains 82% accuracy for those users containing location information in their profile. Roughly, we are able to geolocate correctly around 32% of the users that appear in a monitoring process.

Gender detection is another important factor in many social science studies. A supervised gender classifier is implemented to infer user gender, based on features extracted from academic papers (Kokkos and Tzouramanis, 2014; Rangel et al., 2017). User gender detection is based on classifying messages, no user profile information is used.

10.5 Data Visualization

The GUI has been developed using the Django Web Application framework[20]. This interface provides data analysis visualizations and manages the communication with both the crawler and EliXa.

Talaia implements a number of visualizations which may be customized depending on the needs of the specific use case at hand[21]. The main visualizations include popularity, sympathy and antipathy comparison, evolution of mentions across time, most recent mentions, most widespread mentions, most active users in social media and news sources, and most frequent topics. All those visualizations include interactions that provide further analysis such as looking at the specific data regarding an specific party or candidate, or filtering the data according to various

[19] OpenStreetMap Nominatim (`https://wiki.openstreetmap.org/wiki/Nominatim`) and Google Geocoding API `https://developers.google.com/maps/documentation/geocoding/intro`).

[20] `https://www.djangoproject.com/`

[21] Existing demos and installations implement different visualizations. Visualizations for the use cases described in this paper can be seen at `http://talaia.elhuyar.eus/demo_eae2016` and `http://behagune.elhuyar.eus/`

criteria such as language, time period, data source or author influence. All graph visualizations are implemented using d3.js[22] javascript library.

The interface also has management capabilities which allows to manually review the automatic sentiment labelling. Keyword hierarchy and new website sources can be also set up through the interface. These functionalities ease the process of creating training datasets and adapting Talaia to new domains.

One of the main challenges the interface has to face is how to access data and maintain adequate time responses as the amounts of data gathered escalate. This depends to a great extent on the database optimization, but also on the number of visualizations offered by default. Talaia relies on a Mysql database. The interface does not access data from the actual tables, but rather from a joint view which is refreshed periodically. This reduces the time response from minutes to seconds [23]

10.6 Success Cases

In this section we present two real use cases where Talaia has been applied, and use them for evaluation purposes. The first one focuses on tracking cultural events. The second one analyzes citizen opinions with respect to political parties and candidates during an electoral campaign.

Both monitoring processes presented here ran on their own dedicated servers. We provide details on hardware specifications and volumes of data processed in the following subsections. As a measure of the performance capabilities of our system, the largest monitoring process we have carried out until now gathered 24M tweets per month, with an average of 700K tweets processed per day and a maximum of 1.35M tweets in a single day. Talaia ran on a server with two Intel Xeon 4 core processors (E5530) at 2.4 GHz and 16GB RAM. MySQL databases are locally stored in the server. The crawler and the text analyzers all ran locally, but no interface was implemented in this case.

[22] `https://d3js.org/`

[23] Queries retrieving a million results could take up to 7 minutes (depending on the visualizations required)), while using a joint view takes 40 seconds for the same configuration.

10.6.1 Cultural Domain

Talaia was first applied in the Behagunea[24] project. The objective of the project involved tracking the social media impact of cultural events and projects carried out (more than 500) in the framework of the Donostia European Capital of Culture (DSS2016) year during 2016. The project included monitoring opinions in press and social media in four languages: Basque, French and Spanish as coexisting languages in the different Basque speaking territories and English as international language.

Domain adapted polarity models were created. Since events related to DSS2016 were already programmed during 2015, a previous crawl was carried out in order to build datasets. Those datasets were manually annotated for polarity in a three category scale (positive, negative, neutral). Section 10.7.1 gives more details about the various language and domain specific datasets. Polarity classification models for the cultural domain were trained using those datasets and are distributed as part of EliXa. Section 10.7.2 gives details related to those classifiers.

Talaia ran on an Amazon AWS t4.large dedicated instance[25]. The crawler, the text analyzers and the interface all run locally. A total amount of 166K tweets and press mentions were gathered, with a maximum of 6.6K mentions in a single day. The interface was public and offered real-time results refreshed each 15 minutes. We can see from the volume of the data, that this was a low latency monitoring. Even if there were a lot of events to track, the local nature of most of them explains the little impact they have in social media.

10.6.2 Political Domain

Talaia was used to track citizen opinions during the electoral Basque electoral campaign of September 2016. The crawling was carried out during the election campaign period, starting on September 8th (23:59pm) and finishing on September 23th (23:59pm). It offers useful insights for political analysis such as sympathy rankings, the evolution of the opinions over time, most relevant messages, etc.

The system ran on a server with a Intel Xeon 4 core processor (E5530) at 2.4 GHz and 16GB RAM. MySQL databases are locally stored in the server. The crawler, text analyzers and the interface all run locally. A total amount of 4.25M tweets and press mentions were gathered, with an average of 125K mentions per day, and a maximum

[24] `http://behagune.elhuyar.eus`

[25] Especications are 2 vCPUs, 8GB RAM, 100GB EBS storage disk. More information at `https://aws.amazon.com/ec2/instance-types/`

of 433K mentions in a single day. The interface was public and offered real-time results.

The crawler was configured to find mentions talking about the main political parties present on the campaign and their respective candidates (only main candidates monitored, i.e., those opting to be *Lehendakari*, 'head of the government').

Regarding social media, Twitter was monitored. Since we are talking about monitoring an event happening on a regional scope, two main restrictions were applied: only mentions written in Basque and Spanish were crawled, because those are the two official languages in the region. The second restriction was to constrain mentions to users from the specific geographical area of the Basque Country. The task was then to discard noisy messages, that do not belong to citizens involved in the election, but were likely to be talking about it. In this case, for example the crawling process was likely to capture many mentions from other regions in Spain.

In order to solve this problem we created a census of Twitter users of the Basque Country. We developed a five step algorithm:

(i) We gathered geolocated tweets from the Basque Country area for a certain period of time.

(ii) Authors of those tweets were manually tagged with binary labels, as belonging to the required geographical area or not. Let's call this dataset D_{geo}.

(iii) Taking users tagged as Basque citizens from the previous step, we extended our dataset by retrieving up to the first 5,000 followers and friends from each user using the twitter following API[26]. We compute the frequency of coocurrence for each of the candidates[27], and manually label the most frequent 10,000 candidates. Let's call this dataset $D_{geo+ff-manual}$.

(iv) We train a binary SVM classifier with a linear kernel over $D_{geo+ff-manual}$. Features of the classifier are the number of followers and friends a user has and the relative number of followers and friends (with respect to the total number of follower and friends). The classifier obtains 96% accuracy in a 4-fold cross validation.

(v) Repeat step (ii) with the users labelled as Basque in $D_{geo+ff-manual}$, but this time label the most frequent 20,000 candidates with the classifier trained in step (iv). Our final census consists of 23,195 user ids.

[26] `https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-users-show`

[27] The number of times a candidate appears as a follower or friend of another candidate.

As for the news sources, a list of 30 sources was manually compiled, including TV, printed media and radio stations, all of them with working within the regional scope.

10.7 Evaluation

For the evaluation of Talaia, we evaluate the performance of Elixa's polarity classifier for the two aforementioned domains. In all cases the L2-loss SVM implementation of the LIBLINEAR (Fan et al., 2008) toolkit was used as classification algorithm within Weka (Hall et al., 2009) data mining software. Experiments with polynomial kernels were also conducted (degrees 2-5) but we found no improvement at the expense of much longer training times. All classifiers presented in the following sections were evaluated by 10-fold cross validation. The Complexity parameter was optimized ($C = 0.1$).

For the sake of comparison, all the systems presented from here onwards have been trained using the following set of features:

- 1-gram word forms with frequency >= 2 and document frequency (df) >= 2.
- POS tag 1-gram features.
- Polarity lemmas included in language dependent polarity lexicons. Default lexicons provided with EliXa were used (see Table 10.2 for details).
- Sentence length.
- Upper case ratio: percentage of the capital letters with respect to the total number of characters in a sentence.

Microtext normalization features (URL standardization, OOV normalization and emoticon mapping) are applied before extracting the features of each sentence.

10.7.1 Datasets

Table 10.3 presents the statistics and class distributions of the datasets gathered and annotated in order to build the polarity classifiers for each language in the cultural domain. All annotations were done manually. Polarity was annotated at mention

Language	Lexicon	#neg. entries	#pos. entries	Total entries
eu	$ElhPolar_{eu}$ (San Vicente and Saralegi, 2016)	742	499	1,241
es	$ElhPolar_{es}$ (Saralegi and San Vicente, 2013a)	3,314	1,903	5.217
en	$EliXa_{en}$ (San Vicente et al., 2015)	6,123	3,992	10,115
fr	Feel(Abdaoui et al., 2017)	5.717	8,430	14,147

Table 10.2: Polarity lexicons used in our experiments.

level. Because of the level of specificity reached when defining the keyword taxonomy, we rarely find a mention referring to more than one entity or event. Statistics show that corpora in all languages have a similar distribution, with a high number of neutral mentions, and a larger presence of positive opinions than negative ones.

Language	Total size	#pos	#neg	#neu
eu	2937	931	408	1598
es	4754	1487	1303	1964
en	12,273	4,654	1,837	5,782
fr	11,071	3,459	2,618	4,994

Table 10.3: Multilingual dataset statistics for the cultural domain.

Table 10.4 shows the characteristics of the political domain datasets. In this case, each tweet was annotated with respect to a number of entities appearing in the tweet. Annotators were asked to annotate the polarity of a tweet from the perspective of each of the entities detected in a tweet, that is, a tweet may contain more than one polarity annotation. Example 6 shows a real case where a tweet was given two different annotations, one for each entity (negative expressions underlined, positive ones in bold). In fact, the numbers in table 10.4 give 1.3 and 1.24 average annotations per tweet for Basque and Spanish, respectively.

Example 6

@pnvgasteiz erabat ados, lotsagarria. Aukera ona aurrera begiratu ta ***@ehbildu****—ren euskara arloko proposamena martxan jartzeko #herriakordioa* [28]

Annotating tweets in the political domain proved to be a rather challenging task. Sarcasm is often present, interpellations to a person are frequent even if they are not the target of the opinion, an opinion may be present but in an implicit manner, or a third party negative opinion may be expressed towards an entity but the author may defend it against the expressed opinion. The full annotation guidelines can be consulted in Annex II.

Annotation was carried out in real time during the period of the electoral campaign. We established three shifts a day to annotate messages gathered until then. Three annotators took part in the process. Because of the limited resources and the volume of messages crawled daily, each tweet was annotated by a single annotator.

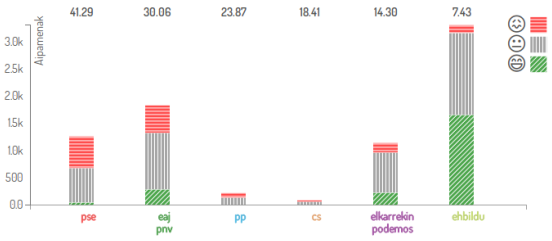


Figure 10.2: Distribution of mentions in Basque with respect to the political parties. From left to right, parties are sorted according the percentage of negative opinions received, with respect to the total amount of mentions received.

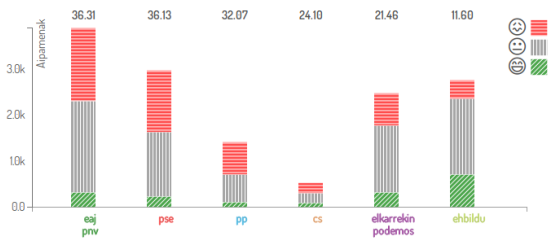


Figure 10.3: Distribution of mentions in Spanish with respect to the political parties. From left to right, parties are sorted according the percentage of negative opinions received, with respect to the total amount of mentions received.

Political domain datasets show very different distributions across languages. While the Basque dataset seems to follow the same pattern seen in the cultural domain, the Spanish dataset has a very high number of negative opinions. Analyzing some result samples, we realized that there are various phenomena that could explain this behaviour. First, much more debate and criticism takes place in Spanish compared to Basque where the tendency is to write a lot more supportive messages.

[28] English translation: @pnvgasteiz totally agrees, shameful. Good chance to look forward and apply the proposal of @ehbildu in the field of Basque #herriakordioa

Those not associated with Basque nationalist ideologies mainly communicate in Spanish. A clear example is that the left party EH Bildu has very few negative mentions in Basque (See figure 10.2). Also the fact that the right wing parties such as *Partido Popular* (PP) and *Ciudadanos* (Cs) receive almost no attention in Basque is a symptom of the little engagement they show in this language.

Second, left wing people were more active in Twitter (in this specific campaign), with right wing parties concentrating the largest amount of negative mentions (see figures 10.3).

Lastly, there is the effect of negative campaigning (Skaperdas and Grofman, 1995), which is more pronounced in Spanish, because as we already said there is much more debate than in Basque. Figure 10.3 also aligns with studies of negative campaigning in multi-party scenarios (Walter, 2014; Haselmayer and Jenny, 2017), being the front-runner *Partido Nacionalista Vasco* (PNV) and its previous government partner *Partido Socialista de Euskadi* (PSE) those who receive the greatest amount of negative mentions.

Language	**#Tweets**	**#Annotations**	**#pos**	**#neg**	**#neu**
eu	9,418	11,692	3,974	3,185	4,533
es	15,550	20,278	3,788	7,601	8,889

Table 10.4: Multilingual dataset statistics for the political domain.

10.7.2 Results

Table 10.5 shows the performance of the various multilingual classifiers trained.

Reported results are in general higher for the cultural domain, even if the datasets are smaller in comparison. Basque and Spanish classifier obtain results above 70%. English and French achieve lower results. Positive mentions present the greatest challenge for English. The main reason for this is the lack of positive training examples. Neutral mentions perform very good in all languages. After analyzing a random sample, we conclude that neutral mentions are of an homogeneous nature, mainly containing agenda events or promotion messages. If we add this to the fact that neutral is the class with the highest number of examples, it seems logical that our classifiers find highly representative features for this class.

Language	#features	acc	fpos	fneg	fneu
Cultural Domain					
eu	4,777	74.02	0.658	0.635	0.803
es	10,037	73.03	0.683	0.756	0.744
en	24,183	70.43	0.715	0.530	0.743
fr	23,779	66.17	0.600	0.617	0.721
Political Domain					
eu	9,394	69.88	0.714	0.702	0.683
es	15,751	67.05	0.545	0.693	0.700

Table 10.5: EliXa polarity classification results.

Regarding the political domain, if we compare Basque and Spanish classifiers, their performance drops around 4% with respect to the results in the cultural domain. Results are not directly comparable, because political domain classifiers are evaluated over entity level tags. Also, political data is more challenging in terms of the linguistic phenomena used. We have detected a fair amount of messages containing sarcasm or opinion ambiguity towards targets.

In the case of Basque, the most sensible drop happens with neutral mentions. After analyzing a random sample we found that they are more heterogeneous than those in the cultural domain. They do contain agenda and promotion messages, but also many third party statements (candidate x says "...") or messages that interpellate parties and candidates over hot topics in the campaign (e.g. "@DanielMaeztu @ehbildu @PodemosEuskadi_ obra gelditzea onuragarria liteke ekonimia arloan 4.000 miloi gastatu eta gero?"[29]). Many neutral messages contain personal opinions not involving any of our predefined target entities, even if they are interpellated. These phenomena make neutral class harder to represent.

Regarding Spanish, performance for positive mentions is significantly lower. Error analysis shows that incorrectly classified instances do not fall into a single category (42% negative, 58% neutral). Analyzing the errors, we find two main reasons. First, many positive mentions are incorrectly classified as negative because their content is mainly negative (e.g. " @AgirreGarita La diferencia es clara, PNV apoyando el desahucio y EHBILDU al desahuciado. NO SEAS COMPLICE, no votes a quien

[29]English translation: @DanielMaeztu @ehbildu @PodemosEuskadi_ stopping the construction would be beneficial after spending 4,000 millions?

desahucia." [30]). Our strategy for assigning message level polarity to all entities involved in a mention is prone to this type of errors. Second, as we saw for Basque, neutral mentions also contain polar expressions or opinions, making it harder to distinguish them from actual positive or negative messages.

10.8 Conclusion and Future Work

We have presented Talaia, a real time monitor of social media and digital press. Talaia is able to extract information related to an specific topic and analyze it by means of natural language processing technologies. Two success cases and the resources generated from those cases have been described. In that sense, we have shown the ability to adapt our system to different domains and languages.

Talaia is still under development. The short term objectives include work on optimizing the information extraction process. Specifically, extracting keywords from the data downloaded up to a certain point would allow us automatically adapt the system to new terms, without losing information because the keyword hierarchy is outdated or the topic is poorly defined.

Another important point is the adaptation of our Sentiment Analysis model to new domains. In that sense experiments are being carried out in order to minimize the domain adaptation process, both in terms of data collection and annotation effort.

Multilinguality is one of the main challenges of such a system. Currently the system is able to process data in 4 languages, and we are working to extend it to new languages.

Furthermore, data analysis may include further processing other than Sentiment Analysis. Geolocation based analysis, user community detection and other useful tasks for user profiling (e.g. gender detection) are the focus of our ongoing work.

All the software behind the platform including the crawler, data processing chain and interface is publicly available under the GNU GPLv3 license.

[30]English translation: @AgirreGarita The difference is that PNV is in favour of evictions and EHBILDU is with the evicted ones. DO NOT BE AND ACCOMPLICE, don't vote to those who practice evictions.

Acknowledgements

This work has been supported by the following projects: Elkarola project (Elkartek grant No. IE-14-382), and Tuner project (MINECO/FEDER grant No. TIN2015-65308-C5-1-R).

Annex I - Comparative of commercial Social Media Monitors

Platform	Data Sources	Crawling	Data Processing	Search	Navigation
Iconoce	Digital press, blogs, videos, social media (Facebook, Twitter, Linked-in?)	Personalized, subject to agreement	no	Personalized archive 3 separate search engines (mentions, comment, authors) no lemmatization - no crosslingual.	Graphs (aggregations?), salient terms, salient topics, Influencers, alerts, reports
INNGUMA	Rss multimedia, Deep Web, Twitter, Facebook, Linkedin, possibility to include external documentation manually	Custom filters, not clear if filtering is done at a post-crawling stage	MT, No mention of text processing. No SA	Semantic search (techniques not especified). Index cards and documents. Information tagged manually.	Reports, content creation, social media management. Multilingual GUI.
Meaning Cloud	Digital news, blogs, Twitter, satisfaction surveys (customer provided), phone survey transcriptions,		5 languages (Es, En, Fr, Pt, It). Language identification, Clustering for topic detection. Lemmatization, pos tagging, parsing, NERC, GATE API SA: Ruled-based. Sentiment Lexicons + rules. Irony and subjectivity detection. Entity polarity detected using manually compiled dictionaries.	no	No Dashboard, visualizations or data aggregatios. Excel plugin or API access
Snap-trends	Social Media (Twitter, Facebook, Instagram, Google+,...)		MT from 80 languages. Propietary linguistic processing. Topic (trends) detection. Propietary sentiment analysis.	Geolocation based search engine, mutiple criteria: social network, search terms, geolocation. Previous search feature.	Agreggations, interactive visualization, temporal trends.

Websays	News, Blogs/RSS, Forums, Facebook, Twitter, Google+, LinkedIn, Instagram, Foursquare, Pinterest, Youtube, Vimeo, Reviews (Tripadvisor, Booking,...)	Keyword based, accepts also negative keywords.	Multilingual data processing, no specific data about the coverage SA: AI (ML) + human validation	Multiple search criteria, filter-based.	Graphs, salient terms, trending topics, influencers, sentiment, trends. Alerts and reports.
Lynguo	Facebook, Twitter, Instagram, YouTube, online media, blogs and forums.	Keyword based, accepts also negative keywords and accounts.	Es,En. SA: ML + Lexicons + rules. Polarity and emotions. Aspect based SA		Customizable dashboard. Several default aggregations and possibility to generate custom visualizations. Alerts and periodical reports
Keyhole	Twitter, Instagram, web sources.	Social media and web sources are configured and monitored separately. Keywords, users.	13 languages. SA: Polarity		Influencers, timeline, trends, sentiment, aggregations.
Uber-metrics	Blogs, forums, academic/scientific journals, digital press, Instagram, Tumblr, Google+, Facebook, Twitter, YouTube, Vimeo, Flickr, and Foursquare. With Ubermetrics you can even capture comments from YouTube, Facebook, and major online news sources. TV/Radio	Customizable "search agents". Keyword based	40 languages. Propietary data processing.	Detailed search based on multiple criteria included in visualization dashboard	Dashboard, alerts, reports.

Table 10.6: Comparison of commercial social media monitoring platforms.

Annex II - Polarity annotation guidelines

We present the guidelines provided to the annotators for marking entity level polarity, including ambiguous cases and the solutions proposed for each of them:

- Neutral: There is no clear opinion or sentiment respect to the target party or candidate from the holder. Mentions referring to objective facts fall into this category as well, even if the fact may be considered positive or negative (e.g. *"El PNV consigue grupo en el senado"*[31]).
- Positive: The mention includes a positive assessment from the holder with respect to the target (e.g. *"Urkullu ha sido un buen lehendakari."*[32]).
- Negative: The mention includes a negative assessment from the holder with respect to the target (e.g. *"Urkullu ha sido un lehendakari mediocre."*[33]).
- ambiguous cases:
    1. Subjectivity is not explicit: *"Cataluña desobedece constantemente la Ley, PNV pide acercamiento de presos, Ribo da los pasos hacia el nacionalismo y Rajoy en SanXenso"*[34]. Main target in the example is "Rajoy" but author expresses a negative opinion towards PNV. Annotators were ask to interpret the implicit subjectivity according to the holder.
    2. The holder expresses the opinion of a third party: *"Podemos cree que Urkullu tiene miedo y por eso adelantará las elecciones - EcoDiario.es <URL>"*[35].The mention expresses a negative opinion from Podemos towards Urkullu. Annotators were asked to annotate it as negative towards Urkullu if they could certify that the holder agreed with the opinion from Podemos, or netural otherways.
    3. There are two (or more) references to a single target, expressing different polarities: *"PNV tendrá grupo propio en el Senado tras la cesión de cuatro*

[31]English translation: PNV gets its own group in the senate

[32]English translation: Urkullu has been a good president

[33]English translation: Urkullu has been a mediocre president

[34]English translation: Catalunya constantly disobeys the law, PNV asks for the rapprochement of prisoners, Ribo makes steps towards nationalism and (meanwhile) Rajoy is in SanXenso.

[35]English translation: Podemos thinks Urkullu is scared and that's why he will call the election early - EcoDiario.es <URL>.

asientos por parte del PP y mantiene su "no a Rajoy" "[36]. The following criteria were applied: N+P=NEU, N+NEU=N, P+NEU=P.

4. The polarity of the message and the polarity towards the target are different: *"El tercer precandidato de #Podemos llama a desalojar al PNV <URL>"*[37]. In those cases, polarity towards the target should be annotated. In the example, message polarity would be neutral, but polarity towards "PNV" would be negative. Thus message would be marked as negative.

5. Irony/sarcasm: *"¿Las cambiamos por Calle Arnaldo Otegi o Paseo de Juana Chaos? Al fin y al cabo, son hombres de paz... <URL>"*[38]. Annotators were asked to interpret irony. The previous example would be thus negative towards the target Arnaldo Otegi.

6. The holder is condemning a negative stance against the target: *"@eldiarionorte cada vez se os ve más el plumero. Panfleto anti bildu. Cuando la salud de los zubietarras empeore, vais y se lo contáis."*[39]. Annotators were asked to interpret the intention of the holder. If the notice a clear intention of defending the target the it should be regarded as positive.

7. The target captured is not the main focus of the opinion: *"@CristinaSegui_ Subió impuestos,no hace nada contra los nacionalistas y les da dinero,no ilegaliza a Bildu y stá implicado en lo de Bárcenas"*[40]. Annotators were asked to mark the polarity towards the target, regardless of the main focus of the opinion.

[36]English translation: PNV will have its own group in the senate thanks to PP handing over four seats, and they still maintain the "No to Rajoy".

[37]English translation: The number three shortlisted candidate of #Podemos calls on the people for throwing PNV out <URL>

[38]English translation: What if we change the name of the street to Arnaldo Otegi St. or Paseo de Juana Chaos? After all, they are men of peace... <URL>

[39]English translation: @eldiarionorte it is increasingly clear what you are up to. Anti Bildu pamphlet. When the people in Zubieta lose their health go and tell them.

[40]English translation: @CristinaSegui_ He increased taxes, he does nothing against nationalists and gives them money, he does not ban Bildu and he is involved in the Bárcenas affair.

PART V

CONCLUSION AND FURTHER WORK

CHAPTER 11

Conclusion and further work

This last chapter presents a summary (Section 11.1) that reviews the objectives posed for this thesis and the goals achieved with respect to Sentiment Analysis (SA) in social media. In Section 11.2 we list the research papers we have published that are related with this work. Section 11.3 describes the software and resources generated. Finally, Section 11.4 proposes some future lines of research.

11.1 Summary

The main goal of this thesis was to research on Multilingual Sentiment Analysis in order to develop a social media monitor on specific topics.

As such, it was important for us that the methods researched be applicable across languages. Because of the presence and importance of the Basque language in our society, special attention has been paid to algorithms which are suitable for less resourced languages.

The first topic addressed in this thesis is the **creation of sentiment lexicons**. Three approaches were analysed:

- Translating existing lexicons in major languages and manually reviewing the sentiment annotations (Saralegi et al., 2013) (chapter 2).

- A corpus-based approach involving the extraction of sentiment bearing terms by means of distributional similarity measures (Saralegi and San Vicente, 2013) (chapters 4 and 8).
- A method based on the propagation of the sentiment of known words through the semantic relations defined in an Lexical Knowledge Base (LKB) such as WordNet (San Vicente et al., 2014) (chapter 3).

Finally, the three aforementioned lexicon construction strategies have been compared in terms of manual effort and performance (San Vicente and Saralegi, 2016) (chapter 4).

The second topic addressed in our research includes the various **challenges that poses analysing messages coming from social media sources**. In a big-data environment, the presence of less resourced languages is insignificant compared to others, non-standard language is used and often several languages are mixed in a single sentence. In relation to this, two main challenges were addressed:

- Language identification was addressed by organizing the TweetLID shared task (Zubiaga et al., 2016) on language identification. As a member of the organizing committee, I took part in the annotation of the datasets, and on the evaluation of the systems (chapter 5).
- Regarding microtext normalization, I took part in the TweetNorm shared task (2013), both as organizer and as participant. As organizer, I took part in the evaluation of the systems, and coordination of the shared task (Alegria et al., 2015) (chapter 6). We also submitted a system as participants (Saralegi and San Vicente, 2013b) which is the basis for the normalization module implemented in EliXa (San Vicente et al., 2015) (chapter 7).

The third main topic of this thesis was **polarity classification**, more specifically, to build classifiers for annotating polarity at sentence and document level. Initially, the knowledge-based classifiers we built relied on average counting of polar words found in sentiment lexicons. However, our main approach was to build a supervised SVM classifier, which allowed us to combine linguistic and statistical features more efficiently. Spanish classifiers won the TASS shared task for in 2012 and 2013 (Saralegi and San Vicente, 2012; Saralegi and San Vicente, 2013) and ranked in the top 3 in 2014 (San Vicente and Saralegi, 2014) (chapter 8). The English classifier was applied

in the Semeval-2015 (Task 12) shared task, achieving notable results in the out of domain testset (San Vicente et al., 2015) (chapter 9). As part of this effort, Basque and French classifiers were also developed (San Vicente et al., 2019) (chapter 10).

Last but not least, and combining all the previous efforts, a real world application has been implemented, a platform that allows real time analysis of the impact of specific topics in social media. Multilingual resources for all the languages mentioned across this thesis are provided in the framework of the project Behagunea. Polarity lexicons, tweet datasets annotated at polarity level, polarity classification models and tweet normalization resources were created for four languages: Basque, English, French and Spanish. The application, composed by three modules (the crawler MSM, the SA tool EliXa and an interface for result visualization) is being further developed as a data analysis product called Talaia[1], and it has already been used successfully in several scenarios (see part IV).

11.1.1 Contributions

The most relevant contributions of this thesis work are listed below:

- We improved the state of the art for Spanish polarity classification, and obtained the first position in the TASS shared task twice (Saralegi and San Vicente, 2012; Saralegi and San Vicente, 2013) and the second position once (San Vicente and Saralegi, 2014) (chapter 8).
- We contributed to the state of the art in aspect based SA for English, and had notable results on the Semeval 2015 aspect based SA shared task (San Vicente et al., 2015) (chapter 9).
- We did pioneering work for Basque in the SA field, specifically:
  - Creating the first sentiment lexicons for Basque (Saralegi et al., 2013; San Vicente and Saralegi, 2016) (chapters 2 and 4).
  - The first polarity annotated datasets for Basque (Saralegi et al., 2013; San Vicente and Saralegi, 2016; San Vicente et al., 2019) (chapters 2, 4 and 10).
  - We generated the first resources for Basque microtext normalization (San Vicente et al., 2019) (chapter 10).

[1] `https://talaia.elhuyar.eus`

- EliXa, The first Multilingual SA system including Basque (San Vicente et al., 2019) (chapter 10).
- Talaia, a real social media monitoring platform applying all the previous research (San Vicente et al., 2019) (chapter 10).

Additionaly, as a result of this research, we distribute a set of robust and open domain tools and resources that are freely available. A detailed list of these resources is presented in Section 11.3.

11.2 Publications

Below, we present chronologically the list of publications related with the research described in this document:

- Xabier Saralegi and Iñaki San Vicente. Tass: Detecting sentiments in spanish tweets. In *Proceedings of the TASS Workshop at SEPLN*, 2012 (chapter 8)
- Xabier Saralegi, Iñaki San Vicente, and Irati Ugarteburu. Cross-lingual projections vs. corpora extracted subjectivity lexicons for less-resourced languages. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, pages 96–108. 2013. ISBN 978-3-642-37255-1 (chapter 2)
- Xabier Saralegi and Iñaki San Vicente. Elhuyar at TASS2013. In *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS2013)*, pages 143–150, Madrid, 2013 (chapter 8)
- Xabier Saralegi and Iñaki San Vicente. Elhuyar at tweetnorm 2013. In *Proceedings of the TweetNorm Workshop at SEPLN*, 2013b (chapter 7)
- Iñaki San Vicente, Rodrigo Agerri, and German Rigau. Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 88–97, 2014 (chapter 3)
- Iñaki San Vicente and Xabier Saralegi. Looking for features for supervised tweet polarity classification. In *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS2014)*, Girona, Spain, September 2014 (chapter 8)

- Iñaki Alegria, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. Tweetnorm: a benchmark for lexical normalization of spanish tweets. *Language Resources and Evaluation*, 49(4):883–905, Dec 2015. ISSN 1574-0218 (chapter 6)

- Arkaitz Zubiaga, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. Tweetlid: a benchmark for tweet language identification. *Language Resources and Evaluation*, 50(4):729–766, Dec 2016. ISSN 1574-0218 (chapter 5)

- Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. Elixa: A modular and flexible absa platform. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 748–752, 2015 (chapter 9)

- Iñaki San Vicente and Xabier Saralegi. Polarity lexicon building: to what extent is the manual effort worth? In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may 2016. ISBN 978-2-9517408-9-1 (chapter 4)

- Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. Real time monitoring of social media and digital press. submitted to Engineering Applications of Artificial Intelligence journal, Elsevier. ISSN: 0952-1976. Preprint available at `https://arxiv.org/abs/1810.00647`, 2019 (chapter 10)

The following references are not included but are very closely related to this thesis:

- Inaki San Vicente and Xabier Saralegi. Polarity classification of tourism reviews in spanish. In *Actas del XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural, SEPLN 2013.*, 2013

- Iñaki San Vicente and Xabier Saralegi. Looking for features for supervised tweet polarity classification. In *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS2014)*, Girona, Spain, September 2014

- Iñaki San Vicente and Xabier Saralegi. Sentimenduen analisirako lexikoen sorkuntza. In *Proceedings of IkerGazte*, 2015

We also list references to other works produced during the development of the present research:

- Iñaki San Vicente, Iñaki Alegria, Cristina España-Bonet, Pablo Gamallo, Hugo Gonçalo Oliveira, Eva Martinez Garcia, Antonio Toral, Arkaitz Zubiaga, and Nora Aranberri. Tweetmt: A parallel microblog corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. ISBN 978-2-9517408-9-1

11.3 Generated resources

11.3.1 Software

- QWN-PPV[2] (chapter 3): QWN-PPV is a method to automatically generate polarity lexicons. It only requires a Lexical Knowledge Base (LKB) such as WordNet and a list of positive and negative seeds (either words or synsets). The algorithm propagates the initial polarities through the LKB projected over a graph by means of the UKB Personalized PageRank algorithm. QWN-PPV is implemented in Java and released under Apache 2.0 license. The distribution includes all needed resources such as graphs representations and dictionaries to generate synset and lemma level annotated polarity lexicons.

- DSPL[3] (Distributional Similarity Polarity lexicons) (chapters 2 and 4): a set of scripts to extract polarity lexicons from corpora, by comparing word co-occurrence information between collections of documents of known similarity. Publicly available under GNU GPL V3 license.

- EliXa[4] (chapter 9): EliXa is a supervised SA system aiming to provide a framework for multilingual aspect based sentiment analysis. Currently it supports training and evaluating sentiment polarity models, and tagging sentence level polarity. The distribution includes resources and sentiment models in four languages. Further details on those resources are given in Section 11.3.4 The framework implements a rule based classifier as well as an SVM based one with a set of features configurable by the user, allowing for experimentation and domain adaptation of the polarity models. Including a new language can be as simple as feeding a polarity lexicon in the correct format to the ruled based classifier. Having further resources such as annotated

[2] `https://github.com/ixa-ehu/qwn-ppv`
[3] `https://github.com/Elhuyar/DSPL`
[4] `https://github.com/Elhuyar/Elixa`

corpora and/or microtext normalization resources can greatly improve the resulting models. The software is written in Java and available under GNU GPL V3 license.

- MSM (Multi Source Monitor)[5] (chapter 10): a keyword-based crawler, which works on a set of keywords defined by the user. MSM is fully integrated in the workflow of Talaia, and thus it benefits from its keyword structures. Rather than a list of unconnected terms, Talaia is designed to work over a hierarchy, which allows a better organization of the data on the analysis step. This way, the keywords are defined as belonging to a specific category. MSM includes a twitter client that connect to the Twitter public stream API, and a syndication feed reader. MSM is available under GNU GPL V3 license.

- Behagunea UI[6] (chapter 10): BehaguneaUI is a Django Web Application that provides data analysis visualizations and manages the communication with both the crawler MSM and the Natural Language Processing (NLP) and SA module EliXa. It is also the interface behind Talaia. It implements a number of visualizations, which may be customized depending on the needs of the specific use case at hand. In addition to data analysis visualizations, the interface offers administrative tools to easily annotate the polarity of the messages gathered, both at message and entity level. All the graph visualizations are implemented using d3.js[7] JavaScript library. Behagunea UI is available under GNU GPL V3 license.

11.3.2 Datasets

- TweetLID dataset[8] (chapter 5): Corpus composed of 35K language annotated messages gathered during march 2014, including tweets from 4 regions in the Iberian Peninsula, where two official languages coexist. The distribution includes the official evaluation script used in the tweetLID shared task.

- TweetNorm dataset[9] (chapter 6: Corpus composed of 1.2K tweets in Spanish where non-standard word forms and their corresponding standardizations

[5] http://github.com/Elhuyar/MSM
[6] https://github.com/Elhuyar/BehaguneUI
[7] https://d3js.org/
[8] http://komunitatea.elhuyar.eus/tweetlid/files/2015/03/TweetLID_corpusV2.zip
[9] http://komunitatea.elhuyar.org/tweet-norm/files/2019/01/tweet-norm_esV3.zip

have been annotated. Messages were gathered in April 2013. The distribution includes the official evaluation script used in the tweetLID shared task.

- Basque Opinion dataset[10] (chapter 4): 200 Polarity annotated Basque sentences belonging to music and film reviews and news articles. The dataset was used for the experiment conducted in San Vicente and Saralegi (2016). It is available under LGPLLR license.

- Behagunea opinion datasets[11] (chapter 10): polarity annotated tweet corpora (Basque,Spanish) about the San Sebastian 2016 cultural capital project generated during San Vicente et al. (2019). Datasets are available under CC-BY-NC-SA license.

- BEC2016 dataset[12] (chapter 10): Dataset including 32K tweet messages in Basque and Spanish languages, harvested during the Basque regional election campaign in September 2016. The tweets were manually annotated with polarity at entity level. The dataset is available under CC-BY-NC-SA license.

11.3.3 Sentiment lexicons

In the following are listed the lexicons generated as result of the research activities carried out during this thesis. Most of these lexicons are publicly available on its own[13] or as part of the EliXa official resources distribution[14].

- Subjectivity lexicons for Basque automatically generated from news paper opinion articles by means of distributional similarity methods, as explained in Saralegi et al. (2013) (chapter 2).

- $ElhuyarPolar_{es}$[15]: Spanish polarity lexicon with manual binary polarity annotations. The lexicon was generated semi automatically from different sources and it is distributed both independently and as part of the EliXa SA software under LGPLLR license. Details about this lexicon can be found in Saralegi and San Vicente (2013); San Vicente and Saralegi (2016) (chapters 4 and 8).

[10] `https://hizkuntzateknologiak.elhuyar.eus/assets/files/basqueopiniondataset-v1.zip`
[11] `https://hizkuntzateknologiak.elhuyar.eus/assets/files/behaguneadss2016-dataset.tgz`
[12] `https://hizkuntzateknologiak.elhuyar.eus/assets/files/bec2016.tgz`
[13] `https://hizkuntzateknologiak.elhuyar.eus/en/resources`
[14] `http://hizkuntzateknologiak.elhuyar.eus/assets/files/elixa-resources-10.tgz`
[15] `http://hizkuntzateknologiak.elhuyar.eus/assets/files/elhpolar-esv1lex.txt`

- $ElhuyarPolar_{eu}$[16]: Basque Polarity lexicon with manual binary polarity annotations. It was generated by translating $ElhuyarPolar_{es}$ and manually reviewing all the resulting polarity annotations. as explained in San Vicente and Saralegi (2016) (chapter 4). Lexicon is available under LGPLLR license.

- QWN-PPV lexicons[17]: English, Spanish and Basque lexicons created by means of the QWN-PPV method, as explained in San Vicente et al. (2014) (chapter 3). They are available under CC-BY-SA V3 license.

11.3.4 Other Resources

- Multilingual Central Repository (MCR) graph representations[18]. These graph representations are used by QWN-PPV to propagate the polarity across MCR synsets.

- Microtext normalization resources for English, Basque, Spanish and French languages[19]. These resources are available under different licenses depending on the ownership of the resource. The distributed package includes the following resources:

  - OOV word and term lists with their respective standard forms. These lists have been compiled from various sources, including corpora-based frequency lists and publicly available web sources.
  - Stopword lists. Note that these are not standard lists containing most frequent words, as they have been compiled for sentiment analysis purposes. Hence, some frequent lemmas are excluded from this lists because they may be relevant to polarity classification (e.g., no, good, ...).
  - Word form dictionaries. These are used as a reference of standard word forms by the microtext normalization module.
  - Emoticon list: this is a list of regular expressions (using Java syntax) matching one or more emoticons, mapped into a 5 category schema, according to the emotion they express. This list is used both at text

[16] `http://hizkuntzateknologiak.elhuyar.eus/assets/files/elhpolar-eullrlex.txt`
[17] `http://adimen.si.ehu.es/web/qwn-ppv`
[18] `http://adimen.si.ehu.es/web/files/qwn-ppv/mcr-graphs_bin.tar.gz`
[19] `http://hizkuntzateknologiak.elhuyar.eus/assets/files/elixa-resources-10.tgz`

normalization step and as clues for polarity classification. The same resource is used for all languages.

- Twitter Sentiment Polarity models trained over cultural domain datasets (San Vicente et al., 2019) for Basque, Spanish, English and French[20].

11.4 Future work

Talaia is the result at production level of the work done in this research. There is however large room for improvement both in the software components and regarding the resources needed to build them.

With respect to sentiment classification, we are directing our efforts towards two lines of work. The first is research for better machine learning algorithms. Deep-learning algorithms have shown great promise in terms of performance for many tasks, including SA. If we take a quick look at the various evaluation campaigns we can see that deep-learning systems are obtaining state of the art results (Nakov et al., 2016; Rosenthal et al., 2017a). We have ongoing experiments for document level polarity classification. Our interests lay on finding an algorithm that maintains a robust performance across domains, but also in measuring the cost of training and hyper-parameter tuning with respect to the improvement obtained over other approaches.

The second line of work is measuring the cost of creating datasets for new domains. Polarity models, and specially those trained on social media suffer significant performance losses when tested over new domain datasets (Pontiki et al., 2015), even when the new domain is close to the original training domain (e.g., hotels vs. restaurants).

With respect to ABSA, we have done a short incursion into the field. The monitoring processes carried out have shown that a more fine grained sentiment analysis is required (San Vicente et al., 2019). In the Basque Election campaign scenario, because we worked in a very restricted domain, we tackled this problem with a predefined set of keywords and entities. Unfortunately, this is not a valid strategy for a more open-ended scenarios, where targets might be unlimited and unknown beforehand (e.g. monitoring tourist activities of a region). The alternative would be to adopt a supervised approach as we did in SemEval (San Vicente et al., 2015), however this means that manually annotated datasets are required for each

[20] `http://hizkuntzateknologiak.elhuyar.eus/assets/files/elixa-models-10.tgz`.

new domain we want to address, which is a major problem. Thanks to shared tasks, some benchmark datasets have been created, but very few domains are covered (i.e., hotels, restaurants, news, electronic products) (Pontiki et al., 2015, 2016; Barnes et al., 2018), and almost all of them are created exclusively for English (except the ABSA for the restaurant domain). One strategy we plan to follow is to generate silver datasets from keyword-based monitoring processes, and later enrich them with entity recognition and manual annotations.

Finally, one of the main issues of a social media monitoring system is to be able to harvest precise data while maintaining a good recall. A crawler such as MSM has the search space limited by the keywords defined. Moreover, the language defining a topic is dynamic, specially in a social media environment where new hashtags are created at every moment. Thus, one of our priorities is to research and implement a *topic detection and tracking (TDT)* solution (Aiello et al., 2013; Osborne et al., 2014). In this sense, in order to extract relevant terms from dynamic document collections, a common approach is to follow an emerging terms detection strategy (Shamma et al., 2011), and then group those terms in clusters defining a topic. Our approach is based on periodically extracting the most salient terms from the data downloaded (Nguyen et al., 2016). Those terms will then be used to update the search space. This would allow us to automatically adapt the system to new terms without losing information (to a certain point) because the the keyword hierarchy is outdated or the topic is poorly defined.

Bibliography

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855, 2017.

Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. Twitcident: fighting fire with information from social web streams. In *Proceedings of the 21st International Conference on World Wide Web*, pages 305–308. ACM, 2012.

Willyan D Abilhoa and Leandro N De Castro. A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, 240:308–325, 2014.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA, 2011. ISBN 978-1-932432-96-1.

Alicia Ageno, Pere R. Comas, Lluís Padró, and Jordi Turmo. The talp-upc approach to tweet-norm 2013. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

Rodrigo Agerri and Ana García-Serrano. Q-WordNet: extracting polarity from WordNet senses. In *Seventh Conference on International Language Resources and Evaluation, Malta. Retrieved May*, volume 25, page 2010, 2010.

Rodrigo Agerri and German Rigau. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82, 2016.

Rodrigo Agerri, Josu Bermudez, and German Rigau. Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, pages 26–31, Reykjavik, Iceland, May 2014.

Aitor González Agirre, Egoitz Laparra, and German Rigau. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *GWC 2012 6th International Global Wordnet Conference*, page 118, 2012.

Eneko Agirre and Aitor Soroa. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, pages 33–41, Athens, Greece, 2009.

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1): 57–84, 2014.

Wasim Ahmed, Peter A Bath, Laura Sbaffi, and Gianluca Demartini. Measuring the effect of public health campaigns on twitter: The case of world autism awareness day. In *International Conference on Information*, pages 10–16. Springer, 2018.

Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, and Alejandro Jaimes. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15 (6):1268–1282, 2013.

Iñaki Alegria, Izaskun Etxeberria, and Gorka Labaka. Una cascada de transductores simples para normalizar tweets. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

Inaki Alegria, Nora Aranberri, Pere R Comas, Vıctor Fresno, Pablo Gamallo, Lluis Padró, Inaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. Tweetnorm_es corpus: an annotated corpus for spanish microtext normalization. In *Proceedings of the Language Resources and Evaluation Conference*, 2014.

Iñaki Alegria, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. Tweetnorm: a benchmark for lexical normalization of spanish tweets. *Language Resources and Evaluation*, 49(4):883–905, Dec 2015. ISSN 1574-0218.

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, Yiming Yang, et al. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop*, volume 1998, pages 194–218. Citeseer, 1998.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on International Language Resources and Evaluation (LREC-2010), Malta.*, volume 25, 2010.

Timothy Baldwin and Marco Lui. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237. Association for Computational Linguistics, 2010.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 127–135, 2008.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Multilingual sentiment and subjectivity analysis. *Multilingual Natural Language Processing*, 2011.

Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA, 2010.

Jeremy Barnes, Toni Badia, and Patrik Lambert. MultiBooked: A Corpus of Basque and Catalan Hotel Reviews Annotated for Aspect-level Sentiment Classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. ISBN 979-10-95546-00-9.

Marco Baroni and Stefano Vegnaduzzo. Identifying subjective adjectives through web-based mutual information. In *Proceedings of the 7th Konferenz zur Verarbeitung Natürlicher Sprache - KONVENS'04*, pages 613–619, 2004.

Eda Baykan, Monika Henzinger, and Ingmar Weber. Web page language identification based on urls. *Proceedings of the VLDB Endowment*, 1(1):176–187, 2008.

Richard Beaufort, Sophie Roekhaut, Louise-Amélie Cougnon, and Cédrick Fairon. A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 770–779, Uppsala, Sweden, 2010.

Kenneth R Beesley. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Proceedings of the 29th Annual Conference of the American Translators Association*, volume 47, page 54. Citeseer, 1988.

Billal Belainine, Alexsandro Fonseca, and Fatiha Sadat. Named entity recognition and hashtag decomposition to improve the classification of tweets. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 102–111, 2016.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. Language identification for creating language-specific twitter collections. In *Workshop on Language in Social Media*, pages 65–74. ACL, 2012.

Dmitriy Bespalov, Bing Bai, Yanjun Qi, and Ali Shokoufandeh. Sentiment classification based on supervised latent n-gram analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 375–382, 2011.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *1010.3003*, October 2010.

Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, 2013.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

Samuel Brody and Nicholas Diakopoulos. Cooooooooooooooollllllllllllll!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 562–570, 2011. ISBN 978-1-937284-11-4.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

Ralf D. Brown. Finding and identifying text in 900+ languages. *Digital Investigation*, 9:S34–S43, 2012.

Ralf D. Brown. Selecting and Weighting NGrams to Identify 1100 Languages. In *Text, Speech, and Dialogue*, pages 475–483, 2013.

Caroline Brun and Claude Roux. Decomposing hashtags to improve tweet polarity classification (décomposition des Â« hash tags Â» pour l'amélioration de la classification en polarité des Â« tweets Â») [in french]. In *Proceedings of TALN 2014 (Volume 2: Short Papers)*, pages 473–478, 2014.

Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*, volume 10, 2010.

MS Cárdenas-Claros and N Isharyanti. Code-switching and code-mixing in internet chatting: Between'yes,"ya,'and'si'-a case study. *The Jalt Call Journal*, 5(3):67–78, 2009.

Simon Carter, Wouter Weerkamp, and Manos Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215, 2013.

Taylor Cassidy, Heng Ji, Lev-Arie Ratinov, Arkaitz Zubiaga, and Hongzhao Huang. Analysis and enhancement of wikification for microblogs with context expansion. In *Proceedings of COLING, the 24th International Conference on Computational Linguistics*, volume 12, pages 441–456, 2012.

William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.

Andrea Ceron, Luigi Curini, and Stefano M Iacus. Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the united states and italy. *Social Science Computer Review*, 33(1):3–20, 2015.

Deepayan Chakrabarti and Kunal Punera. Event summarization using tweets. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, 2011.

Hsinchun Chen, Roger HL Chiang, and Veda C Storey. Business intelligence and analytics: from big data to big impact. *MIS quarterly*, pages 1165–1188, 2012.

Andrey Chepovskiy, Sergey Gusev, and Margarita Kurbatova. Language identification for texts written in transliteration. *CDUD 2012–Concept Discovery in Unstructured Data*, page 13, 2012.

Yejin Choi and Claire Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 590–598, Stroudsburg, PA, USA, 2009. ISBN 978-1-932432-62-6.

Alexander Clark. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 59–66. Association for Computational Linguistics, 2003.

Mathieu Cliche. Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 573–580, 2017.

Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8, 2002.

Marta R Costa-Jussà and Rafael E Banchs. Automatic normalization of short texts by combining statistical and rule-based techniques. *Language Resources and Evaluation*, pages 1–15, 2013.

Juan M. Cotelo-Moya, Fermín L. Cruz, and Jose A. Troyano. Resource-based lexical approach to tweet-norm task. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

Fermín L. Cruz, José A. Troyano, Beatriz Pontes, and F. Javier Ortega. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994, 2014. ISSN 0957-4174.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.

Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3079–3087. 2015.

Amitava Das and Sivaji Bandyopadhyay. Theme detection an exploration of opinion subjectivity. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009*, pages 1–6, September 2009a.

Amitava Das and Sivaji Bandyopadhyay. Subjectivity detection in english and bengali: A CRF-based approach. *Proceeding of ICON*, 2009b.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49, 2015.

Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1045–1052, Republic and Canton of Geneva, Switzerland, 2017. ISBN 978-1-4503-4913-0.

Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on Web search and web data mining*, pages 231–240, 2008.

Gregory Druck. *Generalized expectation criteria for lightly supervised learning*. PhD thesis, University of Massachusetts Amherst, 2011.

Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

Ted Dunning. *Statistical identification of language*. 1994.

Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 359–369, 2013.

Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422, Genoa, Italy, May 2006.

Andrea Esuli and Fabrizio Sebastiani. Pageranking wordnet synsets: An application to opinion mining. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 424–431, Prague, Czech Republic, June 2007.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

Christiane Fellbaum. *WordNet*. 1998.

Christiane Fellbaum and George Miller, editors. *Wordnet: An Electronic Lexical Database*. Cambridge (MA), 1998.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. #hardtoparse: POS tagging and parsing the twitterverse. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, August 2011.

Pablo Gamallo, Marcos Garcia, Susana Sotelo, and José Ramom Pichel. Comparing ranking-based and naive bayes approaches to language detection on tweets. In *TweetLID@SEPLN*, 2014.

Spandana Gella, Kalika Bali, and Monojit Choudhury. "ye word kis lang ka hai bhai?" testing the limits of word level language identification. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 368–377, 2014.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

Moises Goldszmidt, Marc Najork, and Stelios Paparizos. Boot-strapping language identifiers for short colloquial postings. In *Machine Learning and Knowledge Discovery in Databases*, pages 95–111. 2013.

Thomas Gottron and Nedim Lipka. A comparison of language identification approaches on short, query-style texts. In *Advances in information retrieval*, pages 611–614. 2010.

Gregory Grefenstette. Comparing two language identification schemes. In *Proc. of the 3rd International Conference on Statistical Analysis of Textual Data (JADT-95)*, 1995.

Stephen Guo, Ming-Wei Chang, and Emre Kiciman. To link or not to link? a study on end-to-end tweet entity linking. In *HLT-NAACL*, pages 1020–1030, 2013.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, november 2009. ISSN 1931-0145.

Harald Hammarstrom. A FineGrained Model for Language Identification. In *Proceedings of Improving Non English Web Searching (iNEWS'07)*, pages 14–20, 2007.

Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 368–378, 2011.

Bo Han, Paul Cook, and Timothy Baldwin. Lexical normalisation for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 43(1): 15–27, 2013a.

Bo Han, Paul Cook, and Timothy Baldwin. unimelb: Spanish text normalisation. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013b.

Martin Haselmayer and Marcelo Jenny. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & quantity*, 51(6):2623–2646, 2017.

Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, 1997.

Wu He, Shenghua Zha, and Ling Li. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3):464–472, 2013.

Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, 2018.

Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *AAAI'04: Proceedings of the 19th national conference on Artifical intelligence*, pages 755–760, San Jose, California, 2004a. ISBN 0-262-51183-5.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004b.

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. Reconsidering language identification for written language resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2006.

Mans Hulden. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, EACL '09, pages 29–32, 2009.

Mans Hulden and Jerid Francom. Weighted and unweighted transducers for tweet normalization. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

Lluís-F. Hurtado, Ferran Pla, Mayte Giménez, and Emilio Sanchis. Elirf-upv en tweetlid: Identificación del idioma en twitter. In *TweetLID@SEPLN*, 2014.

Norman Ingle. *A language identification table.* 1980.

David Inouye and Jugal K Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *Proceedings of the IEEE Third International Conference on Social Computing (SocialCom)*, pages 298–306. IEEE, 2011.

Laura Jehl, Felix Hieber, and Stefan Riezler. Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 410–421. Association for Computational Linguistics, 2012.

Frederick Jelinek. *Statistical methods for speech recognition*. 1997.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 151–160, 2011.

Valentin Jijkoun and Maarten de Rijke. Bootstrapping subjectivity detection. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 1125–1126, New York, NY, USA, 2011. ISBN 978-1-4503-0757-4.

Valentin Jijkoun and Katja Hofmann. Generating a non-english subjectivity lexicon: relations that matter. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 398–405, 2009.

Yohan Jo and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 815–824, New York, NY, USA, 2011. ISBN 978-1-4503-0493-1.

Rie Johnson and Tong Zhang. Supervised and semi-supervised text categorization using lstm for region embeddings. In *International Conference on Machine Learning*, pages 526–534, 2016.

David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. *ICWSM*, 15:188–197, 2015.

Nobuhiro Kaji and Masaru Kitsuregawa. Automatic construction of polarity-tagged corpus from HTML documents. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 452–459, Stroudsburg, PA, USA, 2006.

Nobuhiro Kaji and Masaru Kitsuregawa. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, EMNLP-CoNLL'07, pages 1075–1083, Prague, Czech Republic, 2007.

Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten De Rijke. Using wordnet to measure semantic orientation of adjectives. In *Proceedings of LREC 2004*, Lisbon, Portugal, 2004.

Max Kaufmann and Jugal Kalita. Syntactic normalization of twitter messages. In *Proceedings of the International Conference on Natural Language Processing*, Kharagpur, India, 2010.

Cynthia Keesan. Identification of written slavic languages. In *Proceedings of the 28th Annual Conference of the American Translators Association*, pages 517–528, 1987.

Gen-itiro Kikui. Identifying, the coding system and language, of on-line documents on the internet. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 652–657. Association for Computational Linguistics, 1996.

Adam Kilgarriff. Comparing corpora. *International journal of corpus linguistics*, 6 (1):97–133, 2001.

Seokhwan Kim, Rafael Banchs, and Haizhou Li. Exploring convolutional and recurrent neural networks in sequential labelling for dialogue topic tracking. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 963–973, 2016.

Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of Coling 2004*, pages 1367–1373, Geneva, Switzerland, Aug 23–Aug 27 2004a.

Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of Coling 2004*, pages 1367–1373, Geneva, Switzerland, August 2004b.

Ben King and Steven P Abney. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, pages 1110–1119, 2013.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. NRC-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland, August 2014.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, 2005.

Athanasios Kokkos and Theodoros Tzouramanis. A robust gender inference model for online social networks and its application to linkedin and twitter. *First Monday*, 19(9), 2014.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the OMG! In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Gustavo Laboreiro, Matko Bošnjak, Luís Sarmento, Eduarda Mendes Rodrigues, and Eugénio Oliveira. Determining language variant in microblog messages. In *Proceedings of the 28th ACM/SIGAPP Symposium On Applied Computing*, pages 902–907. ACM, 2013.

Brian Lehman. The evolution of languages on twitter. `http://blog.gnip.com/twitter-language-visualization/`, 2014.

Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. Twiner: Named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 721–730, 2012.

Bang Hui Lim, Dongyuan Lu, Tao Chen, and Min-Yen Kan. #mytweet via instagram: Exploring user behaviour across multiple social networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 113–120, New York, NY, USA, 2015. ISBN 978-1-4503-3854-7.

Jimmy Lin, Rion Snow, and William Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD)*, pages 422–429. ACM, 2011.

Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. Paraphrasing 4 microblog normalization. In *Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 73–84, 2013.

Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

Fei Liu, Fuliang Weng, and Xiao Jiang. A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1035–1044, Jeju Island, Korea, July 2012.

Hugo Liu and Push Singh. ConceptNet: a practical commonsense reasoning toolkit. *BT Technology Journal*, 22:211–226, 2004.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 359–367, 2011.

Xiaohua Liu, Furu Wei, Shaodian Zhang, and Ming Zhou. Named entity recognition for tweets. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4 (1):3, 2013.

Nikola Ljubešić, Nives Mikelić, and Damir Boras. Language indentification: How to distinguish similar languages? In *Proceedings of the 29th International Conference on Information Technology Interfaces*, pages 541–546. IEEE, 2007.

Marco Lui and Timothy Baldwin. Cross-domain feature selection for language identification. In *In Proceedings of 5th International Joint Conference on Natural Language Processing*. Citeseer, 2011.

Marco Lui and Timothy Baldwin. langid. py: An off-the-shelf language identification tool. In *Proceedings of ACL*, pages 25–30. ACL, 2012.

Marco Lui and Timothy Baldwin. Accurate language identification of twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25, Gothenburg, Sweden, 2014.

Martin Majliš. Yet another language identifier. In *Student Research Workshop at EACL'12*, pages 46–54. ACL, 2012.

Isa Maks and Piek Vossen. Building a fine-grained subjectivity lexicon from a web corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. ISBN 978-2-9517408-7-7.

Bruno Martins and Mário J Silva. Language identification in web pages. In *Proceedings of SAC*, pages 764–768. ACM, 2005.

Paul McNamee. Language identification: A solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3): 94–101, 2005.

Iosu Mendizabal, Jeroni Carandell, and Daniel Horowitz. Tweetsafa: Tweet language identification. In *TweetLID@SEPLN*, 2014.

Zhongchen Miao, Kai Chen, Yi Fang, Jianhua He, Yi Zhou, Wenjun Zhang, and Hongyuan Zha. Cost-effective online trending topic detection and popularity prediction in microblogging. *ACM Transactions on Information Systems (TOIS)*, 35(3):18, 2017.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Annual Meeting of the Association for Computational Linguistics*, volume 45, page 976, 2007.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

Saif Mohammad, Cody Dunne, and Bonnie Dorr. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 599–608, 2009.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. NRC-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, 2013.

Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

Arturo Montejo-Ráez, Manuel Díaz-Galiano, Eugenio Martínez-Cámara, Teresa Martín-Valdivia, Miguel A. García-Cumbreras, and Alfonso Ureña-López. Sinai at twitter-normalization 2013. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

Yerai Doval Mosquera, David Vilares, and Jesus Vilares. Identificación automática del idioma en twitter: Adaptación de identificadores del estado del arte al contexto ibérico. In *TweetLID@SEPLN*, 2014.

Alejandro Mosquera-López and Paloma Moreda. Dlsi en tweet-norm 2013: Normalización de tweets en español. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

Mohamed M Mostafa. More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241–4251, 2013.

Oscar Muñoz-García, Silvia Vázquez Suárez, and Nuria Bel. Exploiting web-based collective knowledge for micropost normalisation. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

Kavi Narayana Murthy and G. Bharadwaja Kumar. Language identification from small text samples. *Journal of Quantitative Linguistics*, 13(1):57–80, January 2006.

Carol Myers-Scotton. *Contact linguistics: Bilingual encounters and grammatical outcomes*. 2002.

Ahmed Nagy and Jeannie Stamberger. Crowd sentiment detection during disasters and crises. In *Proceedings of the 9th International ISCRAM Conference*, pages 1–9, 2012.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. SemEval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June 2013.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, 2016.

Patricia Newman. Foreign language identification: First step in the translation process. Technical report, Sandia National Labs., Albuquerque, NM (USA), 1987.

Dong Nguyen and A Seza Doğruöz. Word level language identification in online multilingual communication. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2014.

Khanh-Ly Nguyen, Byung-Joo Shin, and Seong Joon Yoo. Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information. In *Big Data and Smart Computing (BigComp), 2016 International Conference on*, pages 223–230. IEEE, 2016.

Stefanie Nowak, Hanna Lukashevich, Peter Dunker, and Stefan Rüger. Performance measures for multilabel evaluation: a case study in the area of image classification. In *Proceedings of the international conference on Multimedia information retrieval*, pages 35–44. ACM, 2010.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media*, May 2010.

Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*, 2010.

Jesus Oliva, Jose Ignacio Serrano, M Dolores del Castillo, and Ángel Iglesias. A SMS normalization system integrating multiple grammatical resources. *Natural Language Engineering*, 19(1):121–141, 2013.

Nuno Oliveira, Paulo Cortez, and Nelson Areal. The impact of microblogging data for stock market prediction: using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73:125–144, 2017.

Miles Osborne, Sean Moran, Richard McCreadie, Alexander Von Lunen, Martin Sykora, Elizabeth Cano, Neil Ireson, Craig Macdonald, Iadh Ounis, Yulan He, et al. Real-time detection, tracking, and monitoring of automatically discovered events in social media. *ACL 2014*, page 37, 2014.

Nazan Öztürk and Serkan Ayvaz. Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics*, 35(1):136–147, 2018.

Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

Muntsa Padró and Lluıs Padró. Comparing methods for language identification. *Procesamiento del lenguaje natural*, 33:155–162, 2004.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 271–278, Barcelona, Spain, July 2004.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008. ISSN 1554-0669.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the conference on Empirical methods in natural language processing EMNLP'02 - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002.

John C Paolillo. Conversational codeswitching on usenet and internet relay chat. *Herring, Susan C.(ed.)*, 2011.

W Gerrod Parrott. *Emotions in social psychology: Essential readings*. 2001.

Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. Learning sentiment lexicons in spanish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. ISBN 978-2-9517408-7-7.

Verónica Pérez-Rosas, Carmen Banea, and Rada Mihalcea. Learning sentiment lexicons in spanish. In *LREC*, pages 3077–3081, 2012.

Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. Methodology and construction of the basque wordnet. *Language resources and evaluation*, 45(2): 121–142, 2011.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, 2014.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, 2015.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California, June 2016.

David Pope and Josephine Griffith. An analysis of online twitter sentiment surrounding the european refugee crisis. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, IC3K 2016, pages 299–306, Portugal, 2016. ISBN 978-989-758-203-5.

Jordi Porta. Twitter language identification using rational kernels and its potential application to sociolinguistics. In *TweetLID@SEPLN*, 2014.

Jordi Porta and José Luis Sancho. Word normalization in twitter using finite-state transducers. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

John M Prager. Linguini: Language identification for multilingual documents. In *Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on*, pages 11–pp. IEEE, 1999.

Daniel Preoţiuc-Pietro and Trevor Cohn. A temporal model of text periodicities using gaussian processes. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 977–988, 2013.

Thomas Proisl, Paul Greiner, Stefan Evert, and Besim Kabashi. KLUE: Simple and robust methods for polarity classification. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 395–401, Atlanta, Georgia, USA, June 2013.

Randolph Quirk, Sidney Greenbaum, and Geoffrey Leech. *A comprehensive grammar of the English language*. 1985.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF*, 2017.

Delip Rao and Deepak Ravichandran. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 675–682, Stroudsburg, PA, USA, 2009.

Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*, 2003a.

Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* -, pages 105–112, 2003b.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* -, pages 25–32, Edmonton, Canada, 2003.

Julio Villena Román, Eugenio Martínez Cámara, Janine García Morera, and Salud M Jiménez Zafra. Tass 2014-the challenge of aspect-based sentiment analysis. *Procesamiento del Lenguaje Natural*, 54:61–68, 2015.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval*, volume 14, 2014.

Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, August 2017a.

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, 2017b.

Pablo Ruiz, Montse Cuadros, and Thierry Etchegoyhen. Lexical normalization of spanish tweets with preprocessing rules, domain-specific edit distances, and language models. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

Iñaki San Vicente and Xabier Saralegi. Looking for features for supervised tweet polarity classification. In *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS2014)*, Girona, Spain, September 2014.

Inaki San Vicente and Xabier Saralegi. Polarity classification of tourism reviews in spanish. In *Actas del XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural, SEPLN 2013.*, 2013.

Iñaki San Vicente, Rodrigo Agerri, and German Rigau. Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 88–97, 2014.

Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. Elixa: A modular and flexible absa platform. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 748–752, 2015.

Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. Real time monitoring of social media and digital press. submitted to Engineering Applications of Artificial Intelligence journal, Elsevier. ISSN: 0952-1976. Preprint available at `https://arxiv.org/abs/1810.00647`, 2019.

Iñaki San Vicente and Xabier Saralegi. Polarity classification of tourism reviews in spanish. In *proceedings of "XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural" SEPLN(2013)*, september 2013. ISBN 978-84-695-8349-4.

Iñaki San Vicente and Xabier Saralegi. Sentimenduen analisirako lexikoen sorkuntza. In *Proceedings of IkerGazte*, 2015.

Iñaki San Vicente and Xabier Saralegi. Polarity lexicon building: to what extent is the manual effort worth? In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may 2016. ISBN 978-2-9517408-9-1.

Iñaki San Vicente, Iñaki Alegria, Cristina España-Bonet, Pablo Gamallo, Hugo Gonçalo Oliveira, Eva Martinez Garcia, Antonio Toral, Arkaitz Zubiaga, and Nora Aranberri. Tweetmt: A parallel microblog corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. ISBN 978-2-9517408-9-1.

Xabier Saralegi and Iñaki San Vicente. Tass: Detecting sentiments in spanish tweets. In *Proceedings of the TASS Workshop at SEPLN*, 2012.

Xabier Saralegi and Iñaki San Vicente. Elhuyar at TASS2013. In *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS2013)*, pages 143–150, Madrid, 2013.

Xabier Saralegi and Iñaki San Vicente. Elhuyar at tass 2013. In *Proceedings of the TASS 2013 Workshop at SEPLN*, 2013a.

Xabier Saralegi and Iñaki San Vicente. Elhuyar at tweetnorm 2013. In *Proceedings of the TweetNorm Workshop at SEPLN*, 2013b.

Xabier Saralegi and Iñaki San-Vicente. Elhuyar at tweet-norm 2013. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

Xabier Saralegi, Iñaki San Vicente, and Irati Ugarteburu. Cross-lingual projections vs. corpora extracted subjectivity lexicons for less-resourced languages. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, pages 96–108. 2013. ISBN 978-3-642-37255-1.

Kevin P Scannell. The Crúbadán Project: Corpus building for underresourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, incorporating Cleaneval*, volume 5, page 5, 2007.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

Aliaksei Severyn and Alessandro Moschitti. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 464–469, 2015.

Samira Shaikh, Laurie Beth Feldman, Eliza Barach, and Yousri Marzouki. Tweet sentiment analysis with pronoun choice reveals online community dynamics in response to crisis events. In *Advances in cross-cultural decision making*, pages 345–356. 2017.

David A Shamma, Lyndon Kennedy, and Elizabeth F Churchill. Peaks and persistence: modeling the shape of microblog conversations. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 355–358. ACM, 2011.

Nakatani Shuyo. Language detection library for java, 2010.

Penelope Sibun and Jeffrey C Reynar. Language identification: Examining the issues. In *Proc. of the Fifth Annual Conference on Document Analysis and Information Retrieval (SDAIR), 1996*, 1996.

Penelope Sibun and A Lawrence Spitz. Language determination: Natural language processing from scanned document images. In *Proceedings of the fourth conference on Applied natural language processing*, pages 15–21. Association for Computational Linguistics, 1994.

Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Treviño, and Juan Gordon. Empirical study of machine learning based approach for opinion mining in tweets. In *Advances in Artificial Intelligence*, pages 1–14. 2013.

Vikas Sindhwani and Prem Melville. Document-word co-regularization for semi-supervised sentiment analysis. In *Eighth IEEE International Conference on Data Mining, 2008. ICDM '08*, pages 1025 –1030, December 2008.

Anil Kumar Singh. Study of some distance measures for language and encoding identification. In *Workshop on Linguistic Distances*, pages 63–72. ACL, 2006.

Anil Kumar Singh and Pratya Goyal. A language identification method applied to twitter data. In *TweetLID@SEPLN*, 2014.

Stergios Skaperdas and Bernard Grofman. Modeling negative campaigning. *American Political Science Review*, 89(1):49–61, 1995.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 -*, EMNLP '09, pages 170–179, Stroudsburg, PA, USA, 2009. ISBN 978-1-932432-59-6.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 53–63, Stroudsburg, PA, USA, 2011. ISBN 978-1-937284-13-8.

Andreas Stolcke. Srilm - an extensible language modeling toolkit. In *In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002*, pages 901–904, 2002.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. *The General Inquirer: A Computer Approach to Content Analysis*. 1966.

Carlo Strapparava and Alessandro Valitutti. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Languages Resources and Evaluation (LREC 2004)*, pages 1083–1086, Lisbon, May 2004.

Jeannette N Sutton. Social media monitoring and the democratic national convention: New tasks and emergent processes. *Journal of Homeland Security and Emergency Management*, 6(1), 2009.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2): 267–307, 2011a.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2): 267–307, 2011b.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 133–140, Ann Arbor, Michigan, June 2005.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57. Citeseer, 2013.

Mike Thelwall. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*, pages 119–134. 2017.

Erik Tromp and Mykola Pechenizkiy. Graph-based n-gram language identification on short texts. In *Proc. 20th Machine Learning conference of Belgium and The Netherlands*, pages 27–34, 2011.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July 2010.

Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pages 417–424, Philadelphia, Pennsylvania, 2002.

Peter D Turney and Michael L Littman. Measuring praise and criticism: Inference of semantic oreintation from association. *ACM Transaction on Information Systems*, 21(4):315–346, 2003.

Radim Řehůřek and Milan Kolkus. Language identification on the web: Extending the dictionary method. In *Computational Linguistics and Intelligent Text Processing*, pages 357–368. 2009.

Tommi Vatanen, Jaakko J Väyrynen, and Sami Virpioja. Language identification of short text segments with n-gram models. In *LREC*. Citeseer, 2010.

Jesus Vilares, Miguel A. Alonso, and David Vilares. Prototipado rápido de un sistema de normalización de tuits: Una aproximación léxica. In *Proceedings of the Tweet Normalization Workshop at the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

Julio Villena-Román, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José Carlos González-Cristóbal. Tass-workshop on sentiment analysis at sepln. *Procesamiento del Lenguaje Natural*, 50:37–44, 2012.

Julio Villena Román, Sara Lana Serrano, Eugenio Martínez Cámara, and José Carlos González Cristóbal. TASS-workshop on sentiment analysis at SEPLN. *Proceedings of the Spanish Society for Natural Language Processing (SEPLN)*, 2013.

Julio Villena-Román, Janine García-Morera, Sara Lana-Serrano, and José Carlos González-Cristóbal. Tass 2013 - a second step in reputation analysis in spanish. *Procesamiento del Lenguaje Natural*, 52(0), 2014. ISSN 1989-7553.

John Vogel and David Tresnerkirsch. Robust Language Identification in Short , Noisy Texts : Improvements to LIGA. In *Proceedings of the 3rd International Workshop on Mining Ubiquitous and Social Environments (MUSE)*, pages 1–9, Bristol, UK, 2012.

Annemarie S Walter. Choosing the enemy: Attack behaviour in a multiparty system. *Party Politics*, 20(3):311–323, 2014.

Xiaojun Wan. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 553–561, 2008.

Aobo Wang, Min-Yen Kan, Daniel Andrade, Takashi Onishi, and Kai Ishikawa. Chinese informal word normalization: an experimental study. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP)*, volume 13, pages 127–135, 2013.

Dong Wang and Yang Liu. A cross-corpus study of unsupervised subjectivity identification based on calibrated em. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 161–167, 2011. ISBN 9781937284060.

Xin Wang and Guo-Hong Fu. Chinese subjectivity detection using a sentiment density-based naive bayesian classifier. In *2010 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 6, pages 3299–3304, July 2010.

Zhongyu Wei, Lanjun Zhou, Binyang Li, Kam-Fai Wong, Wei Gao, and Kam-Fai Wong. Exploring tweets normalization and query time sensitivity for twitter search. In *Proceedings of the Text REtrieval Conference (TREC)*, 2011.

Janyce Wiebe, Theresa Wilson, and Matthew Bell. Identifying collocations for recognizing opinions. In *Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, pages 24–31, 2001.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2): 165–210, 2005.

Janyce M Wiebe, Rebecca F Bruce, and Thomas P O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 246–253, 1999. ISBN 1-55860-609-3.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder. In *Proceedings of HLT/EMNLP on Interactive Demonstrations* -, pages 34–35, Vancouver, Canada, 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffman. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, 2005.

Theresa Ann Wilson. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. 2008. ISBN 9780549733249.

Fela Winkelmolen and Viviana Mascardi. Statistical Language Identification of Short Texts. In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*, pages 498–503, Rome, Italy, 2011.

Alexandros Xafopoulos, Constantine Kotropoulos, George Almpanidis, and Ioannis Pitas. Language identification in web documents using discrete hmms. *Pattern recognition*, 37(3):583–594, 2004.

Fei Xia, William D Lewis, and Hoifung Poon. Language id in the context of harvesting language data off the web. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 870–878. Association for Computational Linguistics, 2009.

Jun-Ming Xu, Xiaojin Zhu, and Amy Bellmore. Fast learning for sentiment analysis on bullying. In *Proceedings of the First International Workshop on Issues of*

Sentiment Discovery and Opinion Mining, WISDOM '12, pages 10:1–10:6, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1543-2. doi: 10.1145/2346676.2346686.

Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 129–136, Stroudsburg, PA, USA, 2003.

Ning Yu and Sandra Kübler. Filling the gap: Semi-supervised learning for opinion detection across domains. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 200–209, 2011.

Yang Yu and Xiao Wang. World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans' tweets. *Computers in Human Behavior*, 48:392–400, 2015.

Juglar Díaz Zamora, Adrian Fonseca Bruzón, and Reynier Ortega Bueno. Tweets language identification using feature weighting. In *TweetLID@SEPLN*, 2014.

Marcos Zampieri. Using bag-of-words to distinguish similar languages: How efficient are they? In *Computational Intelligence and Informatics (CINTI), 2013 IEEE 14th International Symposium on*, pages 37–41. IEEE, 2013.

Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. Towards real-time summarization of scheduled events from twitter streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 319–320. ACM, 2012.

Arkaitz Zubiaga, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Vıctor Fresno. Overview of tweetlid: Tweet language identification at sepln 2014. *TweetLID@SEPLN*, 2014.

Arkaitz Zubiaga, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. Tweetlid: a benchmark for tweet language identification. *Language Resources and Evaluation*, 50(4): 729–766, Dec 2016. ISSN 1574-0218.

Arkaitz Zubiaga, Alex Voss, Rob Procter, Maria Liakata, Bo Wang, and Adam Tsakalidis. Towards real-time, country-level location classification of worldwide tweets. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):2053–2066, 2017.