GEPSA, a tool for monitoring social challenges in digital press

Iñaki San Vicente, Xabier Saralegi, Nerea Zubia
Elhuyar Foundation
Zelai Haundi 3,
20170 Usurbil, Spain
{i.sanvicente,x.saralegi,n.zubia}@elhuyar.eus

Abstract

This papers presents a platform for monitoring press narratives with respect to several social challenges, including gender equality, migrations and minority languages. As narratives are encoded in natural language, we have to use natural processing techniques to automate their analysis. Thus, crawled news are processed by means of several NLP modules, including named entity recognition, keyword extraction,document classification for social challenge detection, and sentiment analysis. A Flask 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. Evaluation is provided for the modules related to extraction and classification of information regarding social challenges.

1 Introduction

Today society presents several challenges concerning equality, diversity and inclusion. Among them, this paper focuses on gender equality, migration and minority languages.

When developing policies to address these challenges, evidence-based diagnostics are necessary, and the digital press is a source of evidence or more specifically the narratives of reality that they offer. This type of analysis has already been carried out for several social challenges such as immigration (Chouliaraki and Zaborowski, 2017; Lee, 2019), gender equality (Lansdall-Welfare et al., 2017) and armed conflicts (Khaldarova and Pantti, 2016).

In these narratives we can analyze different aspects or indicators referring to social challenges (importance of the challenge, agents involved, relevant issues in each challenge, etc.). However, narratives are encoded in natural language which implies that to automate their analysis we have to use natural language processing techniques.

In this paper we describe GEPSA, a tool that facilitates the analysis of social challenges on the narratives built in the digital press. For that aim, it automates the analysis of several macro key indicators (KI) associated with each challenge. It also offers navigation to inspect the different indicators at the micro level. Automation is possible thanks to the use of different NLP techniques: entity extraction, keyword extraction, polarity, and text classification. The current demo includes models and resources developed for Basque and Spanish languages.

From here on the paper is organized as follows: the next section gives details about the main features of the system and the indicators extracted for each social challenge analyzed. Section 3 describes the architecture of the tool and its main modules. Section 4 discusses the validation of the tool. The last section summarizes the paper contribution and conclusions drawn from this work.

2 Features

GEPSA is a tool that allows permanent listening of digital press sources. Sources are customizable to each use case, and are defined at the beginning of the monitoring, depending on variables such as the geographical scope or the type of publications to include in the analysis.

Listening is permanent and consists of analyzing different macro indicators for three social challenges: immigration, gender equality, and minority languages. Table1 lists the macro KIs defined for each of the social challenges.

Different filters may be applied interactively to the indicators, thus allowing micro-analysis through access to specific news:

- Time period: Specific time periods may be selected.

Social challenge / indicator	Gender equality	Immigration	Minority Languages
Importance of the challenge on news per source	x	x	x
Importance of the challenge on front-page news per source	x	x	x
Importance of the challenge on news per thematic section		x	x
Gender balance on news	x		
Gender balance among news main characters	x		
Gender balance in front page news	x		
Most relevant persons and organizations in news	x	x	x
Most relevant keywords in news	x	x	x
Most relevant polar words in news	x	x	x

Table 1: Macro key indicators defined for monitoring gender equality, immigration and minority languages.

- Language: System includes news in Basque and Spanish.
- Source: Data corresponding to one or more specific sources may be analyzed.
- Thematic section: News from specific categories (politics, sports, etc.) may be analyzed.

3 System architecture

GEPSA is designed as a pipeline working on a document basis. Its first module is responsible for crawling news data, and retrieved documents are passed through the various modules for information extraction. Each module adds new enriched metadata to a common database.

Figure 1 offers a detailed view of the system components. Lemmatization and Named Entity recognition and classification (NERC) module is responsible for extracting the people and organizations involved in an specific news article, and it also assigns male or female gender to extracted people named entities. Keyword extraction module extracts most salient terms in a document and detects polar words among them. The next module uses supervised classifiers to label the thematic section of a document. The last module detects if a news article belongs to any of the analyzed challenges.

3.1 Crawling

As mentioned earlier, our data source are digital press publications. We mine a set of predefined press sources[1] by means of an in-house RSS feed reader. This reader takes care of gathering news articles periodically[2], removing boilerplate and extracting specific new features (i.e. front page news, photographs).

In order to effectively remove boilerplate from a large number of sources, a hybrid approach has been implemented combining page-level and site-level strategies. In a first step we extract content candidates from the DOM tree using a page-level approach (Kohlschütter et al., 2010), and a second step prunes wrong candidates based on site-level patterns. Those patterns are automatically discovered and updated regularly for each source. Evaluation over 100 random documents (from 20 different sources) shows that the second step improves the number of completely clean documents from 30% to 93%, and benefit is observed for all the types of boilerplate analyzed: banners, breadcrumbs, sharing options, related contents, author/source/date data and photographs.

3.2 NERC

NERC is performed by means of BERT-based (Devlin et al., 2018) neural models. Specifically, BERTeus (Agerri et al., 2020) and BETO (Cañete et al., 2020) models are applied. The Basque model achieves 87.06% Fscore on EIEC corpus (Alegria et al., 2004), while the Spanish model achieves 88.24% Fscore on the CONLL2003 dataset (Tjong Kim Sang and De Meulder, 2003).

In addition to extracting NEs, we detect and classify the gender of person type entities. Detection is done based on gazetteers composed of 6,500 male and 6,500 female names. Those lists were compiled from various sources including international

[1] 180+ sources for the described use case.

[2] Twice a day for the current use case.

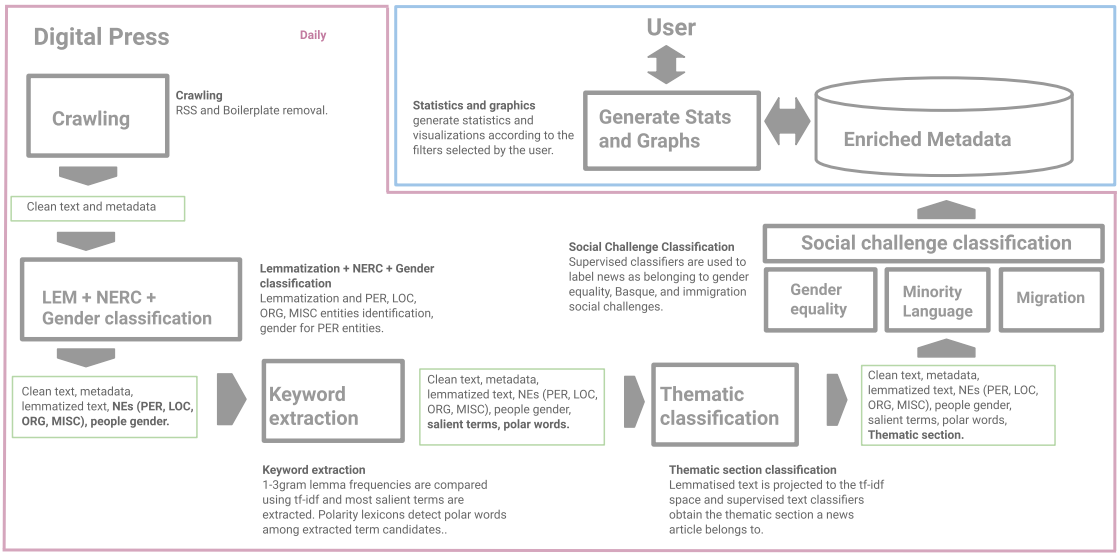

Figure 1: Architecture of GEPSA tool.

and local name lists and news corpora.

3.3 Keyword Extraction

Keyword extraction is done by identifying noun-phrases from the document and computing their representativeness according to TFIDF or Log Likelihood Ratio measures (Dunning, 1993).

For evaluation, keywords were extracted from all the news in annotated challenge datasets (see section 3.5 for details on these datasets), and we manually evaluated the rankings of the most frequent 50 keywords, generated by taking the top n best candidates from each document ($n = [5, 10]$). Precision at different cut offs ($P@[1, 5, 1020, 50]$ was computed. The best performing setup was TFIDF and $n = 5$. P@50 for Spanish was 1 for all social challenges, and ranged from 0.9 to 0.96 for Basque.

Among the keywords extracted, we also detect polar keywords, based on existing polarity lexicons. We used Basque (San Vicente and Saralegi, 2016) and Spanish (Saralegi and San Vicente, 2013) Elh-Polar lexicons.

3.4 Thematic Section Classification

News articles are classified in thematic sections such as politics, culture, sports, etc. There is no standard section taxonomy, press sources have their own criteria and taxonomies. However, it is important to have unified field labels, as this facilitates analyses of the whole narratives depending on the thematic section (e.g. Is it a gender equality diagnostic the same for politics and culture?).

Hence we trained a supervised multi-class classifier in order to classify the crawled news according to a single unique set of eight section classes ([*Economy, Society, Sports, Culture, World, Politics, Technology, Science*]). For that aim we built two news datasets (one per language) by mapping classified news from certain sources to our set of classes. Spanish and Basque datasets contain 156K and 300K classified news, respectively. Systems are trained with 80% of the dataset and evaluated with the rest 20%. Results are reported in table 2. Between Logistic Regression (LR) and Support Vector Machine classifiers (SVM) the former showed better performance, achieving a micro-average F-score of 0.81 and 0.92 for Basque and Spanish, respectively. Representation of news is done by bag-of-words model and TFIDF. Table 2 shows the results by category. Science and Technology classes perform notably worse than the rest due to the lack of examples from those categories in the datasets (less than 1%).

3.5 Social Challenge Classification

Supervised classifiers are used to label news as belonging to gender equality, minority language, and Immigration social challenges. In order to assign a news article to one or more social challenges, supervised binary classifiers were trained. Logistic Regression classifiers are used, using bag-of-words

Category	P	R	F1
Basque			
Economy	0.66	0.84	0.74
Society	0.78	0.68	0.73
Sports	0.92	0.89	0.90
Culture	0.84	0.88	0.86
World	0.80	0.94	0.86
Politics	0.74	0.75	0.74
Technology	0.44	0.87	0.59
Science	0.14	0.24	0.18
micro avg	0.81	0.81	0.81
macro avg	0.66	0.76	0.70
Spanish			
Economy	0.88	0.92	0.90
Society	0.70	0.83	0.76
Sports	0.99	0.98	0.98
Culture	0.93	0.92	0.93
World	0.93	0.92	0.9
Politics	0.93	0.88	0.91
Technology	0.61	0.81	0.69
Science	0.53	0.82	0.64
micro avg	0.92	0.92	0.92
macro avg	0.84	0.84	0.84

Table 2: Precision (P), Recall(R) and F1 score (F1) results for the thematic section classifiers by category.

representations of the documents projected into the TFIDF space. Apart from LR, SVM was tested. LR was chosen over SVM because the latter obtained very low recall values for positive examples of news corresponding to social challenges.

For each social challenge and language, training datasets were created manually. Annotation was carried out by two experts in social research from an specialized consulting service. In order to unify annotation criteria, a first annotation of 100 documents per social challenge was carried out by both annotators, and disagreements were discussed. This resulted in narrowing the scope of some social challenges. From there on each document was annotated once. A minimum of 2,000 documents (per language and challenge) were annotated. In some cases annotation further continued until at least 200 positive examples were found. Datasets were split into 80% for training and 20% for testing. Training sets were oversampled to reach class balance.

Table 3 presents the performance achieved by the classifiers, for each social challenge. Macro averaged measures are reported in this case, because test datasets are highly skewed towards negative examples.

Language	P	R	F1
Basque	0.84	0.80	0.81
Spanish	0.85	0.85	0.85

(a) P, R and Macro-average F-score values achieved by the gender equality classifier on the test sets.

Language	P	R	F1
Basque	0.78	0.83	0.80
Spanish	0.76	0.74	0.75

(b) P, R and Macro-average F-score values achieved by the immigration classifier on the test sets.

Language	P	R	F1
Basque	0.78	0.79	0.78
Spanish	0.83	0.86	0.85

(c) P, R and Macro-average F-score values achieved by the minority language classifier on the test sets.

Table 3: P, R and Macro-average F1 score values achieved by the classifier on the test sets.

3.6 Data Visualization

User can navigate the data using a user interface developed with Flask Web Application framework[3]. GEPSA implements a number of visualizations aimed to represent the indicators defined in section 2. The main visualizations include: news distributions per source and thematic sections; gender balance in overall news, front page news and among protagonists of the articles; most comparison, evolution of mentions across time, most mentioned entities, terms and polar words. All the visualizations include interactions that provide further analysis such as looking at the specific news regarding an specific entity, term or polar word, or filtering the data according to various criteria such as language, time period, data source or thematic section.

4 User validation

A tool such as GEPSA should be validated by users in real case scenarios, if we want such automatic analysis tools to be adopted by social researchers and policy makers. In order to validate GEPSA, a pilot study is being conducted using the tool as diagnostic tool for the aforementioned challenges. There are 5 participants in the study: the two social researchers that participated in the annotation of the datasets, two data scientists and a policy maker. The pilot will last for the first half of 2021 and is

[3] https://palletsprojects.com/p/flask/

expected to generate an intermediate and a final report on aspects such as the validity of the data gathered, the indicators measured, and the usability of the interface. Although we can not yet draw conclusions, preliminary feedback from participants shows that they find the interface intuitive and easy to use, and indicators provide accurate results both for macro and micro analysis.

5 Conclusions

This paper presents a tool that assists researchers analyzing the narratives built by the press respect to specific social challenges, including gender equality, migrations and minority languages.

The developed tool is an example of how by combining various NLP techniques we can automate the process of narrative analysis focused on social challenges.

Moreover, the tool is a showcase of how policy makers may observe those narratives based on evidences collected from a big data environment, according to several predefined key indicators and their temporal evolution. This is useful both to enact more inclusive policies and also to observe the impact of those policies.

Lastly, the user interface and access to data gathered since the beginning of 2020 are publicly available at https://talaia.elhuyar.eus:8978

Acknowledgements

This work has been partially funded by the Basque Government (Gepsa, Hazitek grant no. ZL-2020/00584) and by the Spanish Government (DeepReading, RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE).

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4781–4788.

Inaki Alegria, Olatz Arregi, Irene Balza, Nerea Ezeiza, Izaskun Fernandez, and Ruben Urizar. 2004. Design and development of a named entity recognizer for an agglutinative language. In *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition.*

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020.*

Lilie Chouliaraki and Rafal Zaborowski. 2017. Voice and community in the 2015 refugee crisis: A content analysis of news coverage in eight european countries. *International Communication Gazette*, 79(6-7):613–635.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ted E Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Irina Khaldarova and Mervi Pantti. 2016. Fake news: The narrative battle over the ukrainian conflict. *Journalism practice*, 10(7):891–901.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450.

Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. 2017. Content analysis of 150 years of british periodicals. *Proceedings of the National Academy of Sciences*, 114(4):E457–E465.

Changsoo Lee. 2019. How are 'immigrant workers' represented in korean news reporting?—a text mining approach to critical discourse analysis. *Digital Scholarship in the Humanities*, 34(1):82–99.

Iñaki San Vicente and Xabier Saralegi. 2016. 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).*

Xabier Saralegi and Iñaki San Vicente. 2013. Elhuyar at tass 2013. In *Proceedings of the TASS 2013 Workshop at SEPLN.*

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.