Multilingual and Multimodal Hate Speech Detection in Social Media

Detección Multilingüe y Multimodal de Mensajes de Odio en Redes Sociales

Gretel Liz De la Peña Sarracén
Univeristat Politècnica de València
gredela@posgrado.upv.es

Resumen: In this doctoral thesis we propose the design and development of technologies for the automatic hate speech detection. The hypothesis on which the project is based is that hate detection can be improved by incorporating other sources of information such as images into text processing. In this way, we intend to develop strategies for automatic hate detection from a multimodal approach. Furthermore, the project will take into account the multilingual analysis of hate speech, using transfer learning strategies for the treatment in languages with little information. For the development of the research, we plan to build a dataset that allows multilingual and multimodal processing. In general, the work will be focused on deep learning techniques for the proposal of approaches for hate speech detection.
Palabras clave: Hate Speech Detection, Multilingual System, Multimodal System, Transfer Learning, Deep Learning,

Abstract: En esta tesis doctoral proponemos el diseño y desarrollo de tecnologías para el tratamiento automático de mensajes de odio. La hipótesis en la que se sustenta el proyecto es que la detección de odio puede mejorar al incorporar, en el procesamiento de textos, otras fuentes de información como las imágenes, que en varias ocasiones son compartidas junto a dichos mensajes. De esta forma, pretendemos desarrollar estrategias para la detección automática de odio desde un enfoque multimodal. Por otra parte, en el marco del proyecto tendremos en cuenta el análisis multilingüe de mensajes de odio, haciendo uso de estrategias de transferencia de aprendizaje para el tratamiento en idiomas con poca información. Para el desarrollo de la investigación, nos planteamos construir un conjuno de datos que permita el procesamiento multilingüe y multimodal. En general, el trabajo estará enfocado en técnicas de aprendizaje profundo en la propuesta de aproximaciones para la detección de odio.
Keywords: Detección de Mensajes de Odio, Sistema Multilingüe, Sistema Multimodal, Aprendizaje por Transferencia, Aprendizaje Profundo

CEUR-WS.org/Vol-2802/paper4.pdf

1 Motivation and Background

Nowadays, the web has become one of the main means of communication. Although this is an advantage in many ways, it is also a problem due to the spread of negative messages such as those with hateful content. In general, many sources define a message with hate speech (HS) for any expression whose content is abusive, insulting, intimidating, that incites violence, hatred or discrimination. This type of message is directed against people based on their race, ethnicity, religion, sex, age, physical condition, disability, sexual orientation, political conviction, etc. (Erjavec y Kovačič, 2012; Nobata et al., 2016). Furthermore, these types of messages can contribute to a general climate of intolerance that makes attacks more likely against affected groups[1].

Messages with hate speech are usually viral, so they are potentially dangerous. Therefore, its detection becomes a task of interest in different fields of research. Many of the mechanisms used to identify HS rely on reports from users and content moderators. This is a way that makes it possible to detect

[1] https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech

Proceedings of the Doctoral Symposium on Natural Language Processing from the PLN.net network (PLNnet-DS-2020)
Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)
CEUR Workshop Proceedings (CEUR-WS.org)

some cases of HS, however the large volume of daily publications makes it difficult to detect hate manually. Therefore, building systems capable of detecting HS automatically is a very relevant task.

In fact, automatic hate speech detection has become a trending topic within Natural Language Processing (NLP) in recent years. This has been reflected with the development of different shared tasks related to HS detection. As a result, different research groups have been involved in the treatment of the problem. Thus, approaches based on different strategies for text analysis have been proposed (Schmidt y Wiegand, 2017a; Fortuna y Nunes, 2018; Robinson, Zhang, y Tepper, 2018; Al-Hassan y Al-Dossari, 2019; Zhang y Luo, 2019; Arango, Pérez, y Poblete, 2019; Basile et al., 2019; Schmidt y Wiegand, 2017b). Such approaches range from the use of traditional machine learning techniques to models based on some deep learning architectures.

Some works are based on the use of lexical resources. Others use template-based methods, where words frequently associated with hate expressions and their contexts are identified and subsequently used in detecting hate phrases. On the other hand, traditional statistical methods such as Logistic Regression, Decision Trees and Support Vector Machines have also been widely used in the task. Furthermore, different types of features have been used, such as those obtained with the grammatical labeling (Dinakar, Reichart, y Lieberman, 2011) and the n-grams (Liu y Forss, 2014). Generally, these features are extracted at the lexical level, constructing a vector in which the presence or absence of the words of the defined vocabulary is indicated. Many other works have used methods based on deep learning. In this case, the focus has been more on the architecture of the models than on the linguistic analysis. The most used topologies are convolutional and recurrent networks, and more recently the Transformers models such as BERT. These models have been used independently, but also for the construction of more complex models from the combination of different architectures, as our proposals in some past shared tasks (De la Peña, 2019; De la Peña y Rosso, 2019; De la Peña Sarracén et al., 2018; Muñiz Cuza, De la Peña Sarracén, y Rosso, 2018; De la Peña Sarracén y Rosso, 2019).

Despite the fact that different approaches have been proposed and evaluated for automatic hate detection, the current situation still demands an in-depth analysis in which some factors are taken into account. The motivation of our research arises from the variation of HS given the diversity of languages and the lack of information in the analysis when considering only text. On the one hand, the language can vary due to the diversity of languages. This is that some expressions used in one language to convey hatred do not make sense in other languages when translating it. This is seen even more clearly in variants within the same language. For example, a text in Mexican Spanish, which is not considered hateful in Mexico, may be considered hateful in Spanish from Spain due to the difference in the use of expressions in these countries. Therefore, it is necessary to design strategies that take into account differences between languages and analyze variations within the same language according to different regions when detecting hate. On the other hand, another point to keep in mind is that in hateful texts specific linguistic structures and expressions do not have to appear. In fact, hatred can be expressed implicitly, hence detection can get false results by using only the text as a source. This suggests the use of other sources additional to the text for analysis.

2 Research Hypothesis

The main research hypothesis we want to investigate is that the inclusion of other sources of information besides text to implement multimodal analysis can improve the performance of automatic hate speech detection. In addition, we would like to investigate these topics, including multilingual models that take into account the variability of languages. Therefore, the main objectives of the research can be listed as follows:

i) Design and evaluate multilingual strategies for automatic hate speech detection. We aim to design mechanisms with which different languages can be processed taking into account the variability between them. At the same time, we aim to take advantage of language with a large amount of data to enrich the treatment of other languages with little information by using transfer learning.

ii) Propose multimodal strategies taking into account visual information together with

the textual one. The idea is to incorporate visual content into hate speech detection. We plan to use models based on deep learning designed to process images and text, and combine them for a better understanding and identification of hate.

iii) Create a dataset containing different languages and information sources. As part of this objective, we aim to design a mechanism to obtain phrases that can express hatred to build linguistic resources for the construction of the corpus. We would like to propose a semi-supervised strategy to tag the corpus that will be multimodal by containing text, images, and geolocation information.

iv) Experiments on the new dataset with various approaches on the system and features. With the new dataset we pretend to carry out experiments to evaluate the previously used approaches. In addition, we plan to design new approaches considering the new geolocation information that will allow us to study the publication of hate speech in geographic areas.

3 Methodology and Experiments

According to the objectives, we started by participating in two shared tasks. We proposed and evaluated multimodal models in one of the tasks and a multilingual system in the other. Relevant details are described below:

3.1 Memotion Analysis (SemEval-2020 Task 8)

Memotion Analysis shared task has been organized in order to bring the attention towards Internet memes processing (Sharma et al., 2020). Three subtasks were defined in the task, which are Sentiment Classification (subtask A), Humour Classification (subtask B) and Scales of Semantic Classes (subtask C). The main goal of the subtask A is to classify a meme as positive, negative or neutral according to sentiment content. The subtask B aims to identify the type of humour expressed among the categories: sarcasm, humour, offense and motivation. These concepts are closely related to hateful content. In particular, sarcasm is a phenomenon that must be considered in hate speech detection, since hate is often transmitted implicitly through sarcasm. Finally, the subtask C focuses on quantifying the level to which a particular category is expressed.

We proposed a multilingual model that combines the analysis of textual and visual information (De la Peña y Rosso, 2020c). The pretrained BERT base model is used for the text processing, and a pretrained VGG model for images. Features obtained from both text and image analysis are combined to feed a simple classifier that obtains the final categories, as Figure 1 shows. We perform an ablation study in course of the proposal design to analyze the importance of the multimodal model. Furthermore, we analyze different models rather than BERT and VGG, including traditional machine learning models such as Support Vector Machines.

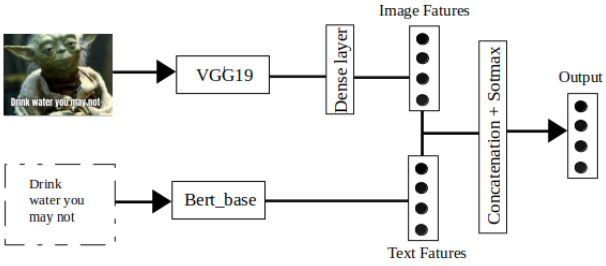


Figure 1: Multimodal model

The image features vector is concatenated with the textual features vector to obtain a general features vector. This vector is a high level representation of a meme and is used in the final softmax layer to obtain the output.

In the experiments, we used macro F1 for each of the subtasks, and then the average. Table 1 shows a comparison of the average among the results obtained with the proposed model and variants in which some of its components are removed. Basically, in each evaluated model, one of the models that processes text or images was eliminated. It should be noted that when a simple model is used instead of the multimodal model, the results are even worse.

Model	Average
Multimodal Model	**0.4300**
Text Model	0.3042
Image Model	0.3306

Table 1: Macro F1 results

Table 2 shows the results obtained in the test set for subtasks B and C, which are related to hate speech detection. The number of participants was 31 and 28 for the subtasks respectively, where we reached the positions 4

and 6. When analyzing the general ranking, it can be seen that even the best results were around to the value 0.5, which suggests once again that the models do not learn correctly. We believe that the values are due to the labelled of the memes, since for each category is not very clear the annotation that can be very subjective. The quality of the dataset provided in this task increases the need for the creation of the multimodal corpus that we consider as part of the doctoral thesis.

Model	Subtask B		Subtask C	
	Pos	F1	Pos	F1
Best system	1	0.5183	1	0.3224
Our system	4	0.5093	6	0.3143
Last system	32	0.4002	29	0.1267
Baseline	10	0.5002	19	0.3008

Table 2: Results in the test set

3.2 Multilingual Offensive Language Identification in Social Media (SemEval-2020 Task 12)

Multilingual Offensive Language Identification in Social Media is a shared task which aims to identify offensive language in texts. The languages included in the task are English, Arabic, Danish, Greek and Turkish (Zampieri et al., 2020).

We propose a system based on BERT (De la Peña y Rosso, 2020b). The first step in the strategy is a preprocessing. In this step the English tweets are cleaned. Firstly, misspelled words are corrected with the support of the *TextBlob*[2] tool. We think it is an important step since many users tend to misspell words and this can lead to a large number of elements outside of the vocabulary. Also, we replaced each emoji with a phrase that describes its meaning with *emoji*[3] tool.

3.2.1 Features

We included a feature analysis to use the information for discrimination between classes. The first group of features is based on some texts *basic properties*: (i) the length of the tweets, (ii) the number of misspelled words and iii) the use of punctuation marks, which is the number of times that one of the signs in the set {?! '[...]} is used in the text, which indicate exclamation, question or omission of phrases. The element [...] corresponds to a sequence of more than one dot. Another group of features is based on *semantic properties*: (i) the use of emoticons, as well as (ii) the noun phrases. In both cases, a vector is constructed with the emoticons or noun phrases present in a text. The representation in this vector space is based on TF-IDF and the dimensionality of the vectors is reduced by using the Principal Component Analysis (PCA) technique. Then, it is added to this vector a last component indicating the number of emoticons or noun phrases in the original text.

[2] `https://textblob.readthedocs.io/en/dev/`

[3] `https://github.com/carpedm20/emoji/`

3.3 Method

The general architecture of the proposed system is showed in Figure 2.

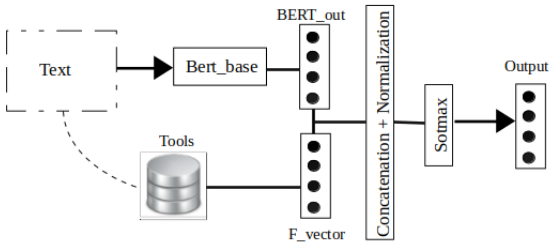


Figure 2: General system architecture

The system consists of a BERT based model at the text level. This model is used as an embedding generator from the text. Hence, a vector representation (BERT_out) is obtained given a text. Basically, the vector is the output of the special token [CLS] included in the processing in BERT. Afterward, this vector is concatenated with the features vector obtained before, and a normalization layer is applied to the result. Finally, the vector is fed to a softmax layer to predict the output.

In the analysis of the Arabic, Danish, Greek and Turkish languages, the same architecture was used. The main difference lies in the model used to obtain the vector of text embeddings. In this case the MBERT model is used, which has been trained with 104 languages in the same way as BERT for English. In the model the tokens from different languages share an embedding space and a single encoder. There are no cross-lingual objectives specifically designed nor any cross-lingual data, like parallel corpora. However, MBERT produces a representation that seems to generalize well a cross languages for

a variety of tasks.

The experiments are carried out with the 10-fold cross validation stratified technique. The measure is macro F1-score, according to the one used for the ranking of the systems in the competition.

Table 3 shows a summarization of the experimental results for offensive language identification, obtained for English. In the proposed model (Proposal) all the features are taken into account and BERT is used. We can check the superiority of the proposal compared to the baselines.

Proposal	**0.9496**
Baselines	
SVM	0.8825
CNN	0.8910
BiLSTM	0.9204

Table 3: Macro F1 for English

Table 4 shows the results for languages other than English. On the one hand, we can see that the features are not very relevant, since the difference in the results is not significant with respect to those obtained with the model where the features are not used. For Arabic, Greek and Turkish, the proposed model achieves better results with respect to the baselines as well as in English.

Model	**Languages**			
	Arabic	**Danish**	**Greek**	**Turkish**
Proposal	**0.8064**	0.7048	**0.7350**	**0.7218**
Baselines				
SVM	0.6912	**0.7258**	0.7295	0.7033
CNN	0.7343	0.6503	0.6782	0.6675
BiLSTM	0.7556	0.6620	0.7049	0.6932

Table 4: Macro F1 for Arabic, Danish, Greek and Turkish

3.4 Keyword Extraction and BERT-based Transfer Learning

The current work is focused on proposing a strategies to extract words and phrases that can express hatred. We study different approaches that have been proposed for keyword extraction in a general context. Then, we design a strategy that uses tagged corpus to extract phrases taking into account not only the frequency of appearance in hateful texts, but also analyzing the non-hateful texts of the corpus. An interesting issue is that some users often use words that can indicate hate, in contexts where hate is not actually expressed.

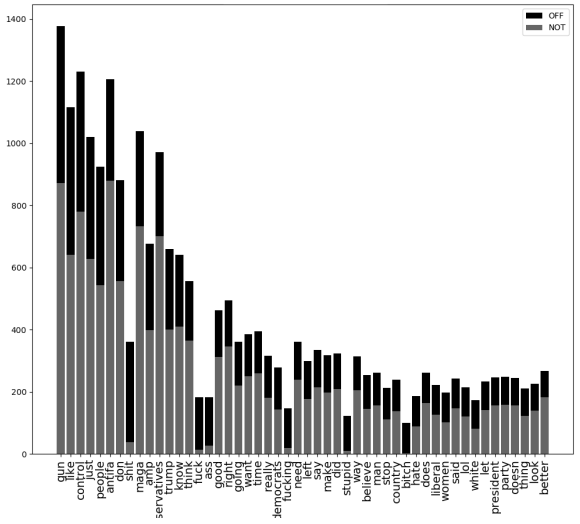
According to this issue, in (De la Peña y Rosso, 2020a) we propose a strategy which uses a new measure to rank offensive words from a tagged corpus on the basis of the cumulative distribution function and the harmonic mean of relative frequencies of the terms of offensive and non-offensive texts.

We use two scores based on relative frequencies and the Cumulative Distribution Function. So that, we obtain two measures for each words that describe the terms distribution in a cumulative way. Finally, the harmonic mean is used to combine both measures. It gives the greatest weight to the smallest item of a series, so that it gives a curve straighter than the arithmetic one. Thus, we obtain the final measure.

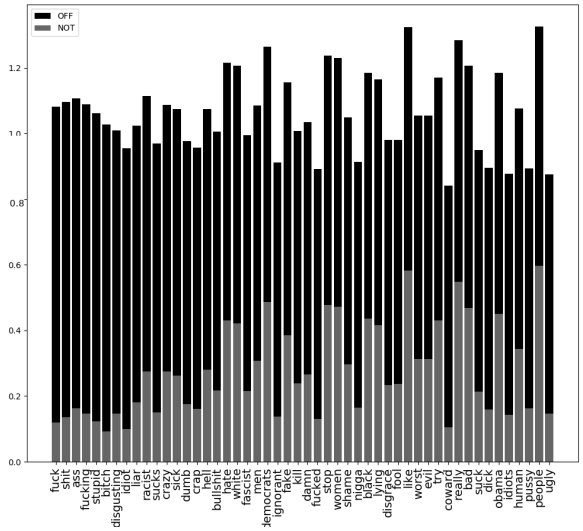
Figures 3(a) and 3(b) show the 50 most relevant terms according to the frequency and the new measures for the offensive class respectively. In the analysis we use the OffensEval 2019 data set (Zampieri et al., 2019). The darker portion of each bar indicates the measure value for the term in the offensive class and the lighter part for the non-offensive class. We can check that the high frequency of a term in offensive texts is not a good discriminator. The first terms in the ranking also have a high frequency in non-offensive texts. The frequency is even greater for the non-offensive class than for OFF in most cases. In contrast, when using the new measure the first terms like 'f*ck', 'sh*t' and 'a*s' are more discriminative.

Furthermore, we proposed an approach based on transfer learning for the multilingual offensive language detection. The model uses MBERT to represent mapping between languages, which is concatenated with other BERT-based model, trained for offensive language identification in English. Then, we include in the model a residual connection from the MBERT to the classifier output, followed by a layer normalization. The idea is to directly add the corresponding information from the generated representation of the text to the information obtained from the pre-trained model for classification. Finally, the complete model is trained for each new language.

We evaluated this proposal for the four languages in the dataset of Offenseval 2020,



(a) Frequency



(b) New Measure

Figure 3: Most relevant offensive terms

taking English as the original language. In all languages the results are better when the transfer learning strategy (among languages) is applied, taking as reference the results obtained when we used only MBERT. Also, we analyzed the residual connection of the model, and the results get worse when it is eliminated.

4 Future Work

Once the shared task have finished and results are obtained to extract offensive words, we plan to begin the construction of the multilingual and multimodal dataset. Therefore, we will follow the steps listed below:

1. Filter messages on the Twitter social network using phrases taken from the previous process and words taken from the Hatebase[4] tool. We will use the Twitter API to download tweets in a certain period of time.

2. Find potential haters by analyzing the users who publish the downloaded tweets and already identified accounts on Twitter.

3. Download the tweets of the identified users in real time to obtain the publications that have not yet passed the Twitter filter.

   In this step tweets will be filtered for the identified users, downloading those that contain, in addition to the text, visual and geolocation information.

4. Develop experiments with different models on the new built dataset.

   The idea is to evaluate models based on deep learning with architectures such as recurrent and convolutional networks, as well as Transformer models like BERT and ALBERT. On the other hand, we propose to design and evaluate strategies based on deep representation learning techniques on knowledge graphs analysis. Our purpose is to incorporate multimodal information and to address the explainability in the hate speech detection with the analysis on graphs.

Bibliografía

[Al-Hassan y Al-Dossari2019] Al-Hassan, A. y H. Al-Dossari. 2019. Detection of Hate Speech in Social Networks: A Survey on Multilingual Corpus. En *6th International Conference on Computer Science and Information Technology*.

[Arango, Pérez, y Poblete2019] Arango, A., J. Pérez, y B. Poblete. 2019. Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation. En *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 45–54.

[Basile et al.2019] Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, y M. Sanguinetti. 2019. Semeval-2019 task 5: Multilingual Detection of Hate Speech against Immigrants

[4] `https://hatebase.org/`

and Women in Twitter. En *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 54–63.

[De la Peña2019] De la Peña, G. L. 2019. GL at SemEval-2019 Task 5: Identifying Hateful Tweets with a Deep Learning Approach. En *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 416–419.

[De la Peña y Rosso2019] De la Peña, G. L. y P. Rosso. 2019. DeepAnalyzer at SemEval-2019 Task 6: A Deep Learning-based Ensemble Method for Identifying Offensive Tweets. En *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 582–586.

[De la Peña y Rosso2020a] De la Peña, G. L. y P. Rosso. 2020a. Harmonic Mean of Relative Frequencies of Terms and BERT-based Transfer Learning for Multilingual Offensive Language Detection. *Submitted*.

[De la Peña y Rosso2020b] De la Peña, G. L. y P. Rosso. 2020b. PRHLT-UPV at SemEval-2020 Task 12: BERT for Multilingual Offensive Language Detection. *Accepted at SemEval 2020*.

[De la Peña y Rosso2020c] De la Peña, G. L. y P. Rosso. 2020c. PRHLT-UPV at SemEval-2020 Task 8: Study of Multimodal Techniques for Memes Analysis. *Accepted at SemEval 2020*.

[De la Peña Sarracén et al.2018] De la Peña Sarracén, G. L., R. G. Pons, C. E. Muñiz-Cuza, y P. Rosso. 2018. Hate Speech Detection Using Attention-based LSTM. En *EVALITA@ CLiC-it*.

[De la Peña Sarracén y Rosso2019] De la Peña Sarracén, G. L. y P. Rosso. 2019. Aggressive Analysis in Twitter using a Combination of Models. En *In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings*.

[Dinakar, Reichart, y Lieberman2011] Dinakar, K., R. Reichart, y H. Lieberman. 2011. Modeling the Detection of Textual Cyberbullying. En *Fifth International AAAI Conference on Weblogs and Social Media*.

[Erjavec y Kovačič2012] Erjavec, K. y M. P. Kovačič. 2012. "You Don't Understand, This is a New War!" Analysis of Hate Speech in News Web Sites' Comments. *Mass Communication and Society*, 15(6):899–920.

[Fortuna y Nunes2018] Fortuna, P. y S. Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

[Liu y Forss2014] Liu, S. y T. Forss. 2014. Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification. En *KDIR*, páginas 530–537.

[Muñiz Cuza, De la Peña Sarracén, y Rosso2018] Muñiz Cuza, C. E., G. L. De la Peña Sarracén, y P. Rosso. 2018. Attention Mechanism for Aggressive Detection. En *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, (IberEval 2018), CEUR Workshop Proceedings. Vol. 2150. Pages 114-118.*

[Nobata et al.2016] Nobata, C., J. Tetreault, A. Thomas, Y. Mehdad, y Y. Chang. 2016. Abusive Language Detection in Online User Content. En *Proceedings of the 25th International Conference on World Wide Web*, páginas 145–153. International World Wide Web Conferences Steering Committee.

[Robinson, Zhang, y Tepper2018] Robinson, D., Z. Zhang, y J. Tepper. 2018. Hate Speech Detection on Twitter: Feature Engineering vs Feature Selection. En *European Semantic Web Conference*, páginas 46–49. Springer.

[Schmidt y Wiegand2017a] Schmidt, A. y M. Wiegand. 2017a. A Survey on Hate Speech Detection using Natural Language Processing. páginas 1–10, 01.

[Schmidt y Wiegand2017b] Schmidt, A. y M. Wiegand. 2017b. A Survey on Hate Speech Detection using Natural Language Processing. En *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, páginas 1–10.

[Sharma et al.2020] Sharma, C., W. Paka, Scott, D. Bhageria, A. Das, S. Poria, T. Chakraborty, y B. Gambäck. 2020. Task Report: Memotion Analysis

1.0 @SemEval 2020: The Visuo-Lingual Metaphor! En *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, Sep. Association for Computational Linguistics.

[Zampieri et al.2019] Zampieri, M., S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, y R. Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, páginas 1415–1420.

[Zampieri et al.2020] Zampieri, M., P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, y c. Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). En *Proceedings of SemEval*.

[Zhang y Luo2019] Zhang, Z. y L. Luo. 2019. Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. *Semantic Web*, 10(5):925–945.