GSI-UPM at SemEval-2019 Task 5: Semantic Similarity and Word Embeddings for Multilingual Detection of Hate Speech Against Immigrants and Women on Twitter

Diego Benito, Oscar Araque, and Carlos A. Iglesias
Intelligent Systems Group
Universidad Politécnica de Madrid
Madrid, Spain
Avenida Complutense, 30
{d.benito,o.araque,carlosangel.iglesias}@upm.es

Abstract

This paper describes the GSI-UPM system for SemEval-2019 Task 5, which tackles multilingual detection of hate speech on Twitter. The main contribution of the paper is the use of a method based on word embeddings and semantic similarity combined with traditional paradigms, such as n-grams, TF-IDF and POS. This combination of several features is fine-tuned through ablation tests, demonstrating the usefulness of different features. While our approach outperforms baseline classifiers on different sub-tasks, the best of our submitted runs reached the 5th position on the Spanish subtask A.

1 Introduction

Information available in social networks is the result of many interactions between users and their activity on the net. Unfortunately, hate speech and other misuses are proliferating on the Internet. Hate speech authors justify their conduct based on the freedom of speech argument. Thus, a debate over hate speech legislation and freedom of speech has been generated (Herz and Molnar, 2012).

The task to decide if a piece of text contains hate speech is not trivial, even for humans. Being subject to different interpretations and opinions, the manifestations of hate speech become difficult to define. Based on previous hate speech statements (Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017), this phenomenon could be defined as offensive or humorist content in form of text, video, or images that attacks, diminishes, incites violence or hate against groups or individuals, based on actual or perceived specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, or any other.

Hate speech topic has gained impact and popularity in recent years, which is reflected not only by the increased media coverage but also by the growing political attention. Regarding the specific forms of hate speech that we deal with, sexism and racism victims increased during 2017 according to the FBI hate crime statistics[1].

For this reason, participating in SemEval2019 Task 5 (Basile et al., 2019) is such an interesting challenge. The proposed task consists in Hate Speech detection in Twitter messages featured by two specific different targets intrinsically related to the phenomena mentioned above, immigrants and women. The task is enriched by adding a multilingual perspective fostering the research for both English and Spanish messages.

The system proposed relies on a supervised classifier using different text features combined with several strategies with the aim of finding an optimal performance. The remainder of this paper is structured as follows. After this introductory section, Section 2 reviews related work. Following, the proposed classification model is described in Section 3. Then, Section 4 presents the experimental results, and finally, Section 5 concludes the paper with a final discussion.

2 Related Work

Most of our literature review from the field is referenced by previous survey research (Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017).

Multiple procedures have been implemented, since more traditional feature engineering, such us n-grams (Waseem and Hovy, 2016) or Part-of-Speech (POS) (Davidson et al., 2017) to more complex deep learning architectures (Yuan et al., 2016; Badjatiya et al., 2017).

According to the analyzed bias which motivates hate speech, general hate speech (Silva et al., 2016) is considered by the majority, however,

[1] https://ucr.fbi.gov/hate-crime/

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 396–403
Minneapolis, Minnesota, USA, June 6–7, 2019.

there is large research that focuses particularly on racism (Kwok and Wang, 2013) and sexism (Hewitt et al., 2016). Though it is not exactly a form of hate speech, cyberbullying is a very related problem with some study research (Cortis and Handschuh, 2015).

3 System Overview

The system relies on a supervised machine learning algorithm. This final classification step is fed by a data processing pipeline formed by the preprocessing and the feature extraction modules. Regarding the implementation, Python has been used, with the additional capabilities provided by the libraries scikit-learn (Pedregosa et al., 2011), NLTK (Bird and Loper, 2004), and GSITK (Araque et al., 2017) [2]. Figure 1 illustrates the system architecture from a general perspective.

3.1 Preprocessing

In this phase, the raw text is taken and cleaned using common NLP techniques (Manning et al., 1999): removal of punctuation marks, special characters, URLs, and stop-words. Tweet preprocessing relies on tokenization, user mentions normalization, the appearance of hashtags, URLs, and all caps words flagged supported by the tools provided by GSITK. In addition, tokens are lemmatized using the Porter stemmer (Porter, 1980).

3.2 Feature Engineering

Different features have been taken into account during the feature engineering stage. Such features are divided into subcategories: statistical features, content analysis, word embeddings, semantic features, and linguistic features.

3.2.1 Statistical Features

We collected word and character n-grams evaluating both approaches, Bag-of-Words (BOW) and Term Frequency - Inverse Document Frequency (TF-IDF). The reason to include character n-grams comes from the Twitter domain, where texts are short and misspelling may occur; this can be attenuated at the character level (Schmidt and Wiegand, 2017). Apart from the mentioned reasoning, previous research (Mehdad and Tetreault, 2016) has shown the effectiveness of character n-grams in the problem of offensive language.

Besides tokens included within the text corpus, the system also includes frequencies from external lexicons that are thought for hate speech[3], sentiment analysis (Hu and Liu, 2004; Liu et al., 2005), and subjectivity analysis (Pang and Lee, 2004).

3.2.2 Content Analysis

As seen, sentiment and subjectivity information has been included. Hate speech can be considered as subjective content, and a relation between subjectivity, sentiments, and emotions can occur. Besides, hate speech is expected to have a negative polarity, so text subjectivity and polarity provided by the TextBlob (Loria et al., 2014) library were included in the analysis.

Topic modeling methods were added to the study, particularly, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) in order to extract the topic of each tweet in combination with the appearance of hashtags (topics) inside the corpus.

3.2.3 Word Embeddings

In order to solve the lack of semantic of words in n-grams features, word distributed representations based on word embeddings models are evaluated. Pre-trained word vectors convert words into vector space where semantically similar words tend to appear close by each other. In this system, a vector is extracted for each word in the input text; then, as done in (Araque et al., 2017), the average pooling operation is performed on all word vectors, resulting in a vector of the same dimensions as the original word vectors.

3.2.4 Semantic Features

A central part of the system consists of a method (Araque et al., 2019) that exploits the semantic similarity measure that a word embedding model provides, via cosine similarity. In general lines, this approach uses a lexicon to which the input text is projected, employing the similarity measure obtained from an embedding model.

The method considers a selection of words S that constitutes a lexicon vocabulary to which the input documents are projected. Given a text document (e.g., tweet), a similarity value between the input word vectors of that document and each of the words in S is computed. After iterating over all input words and all lexicon words, a matrix $m \times |S|$ is obtained, where m is the number of input words in a particular document. Following,

[2] https://github.com/gsi-upm/semeval2019-hatespeech

[3] https://hatebase.org/

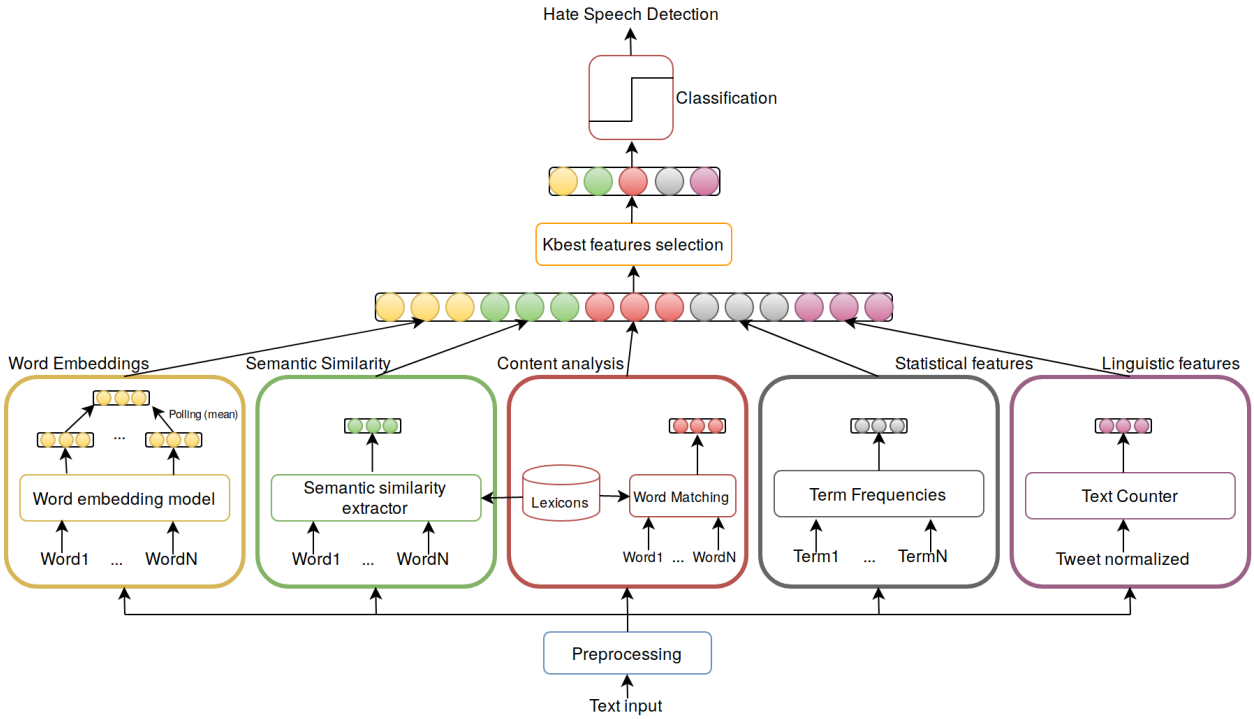

Figure 1: System Overview

the maximum pooling function is applied column-wise, obtaining the semantic similarity feature vector of dimensionality $|S|$.

In this work, the previously mentioned lexicons have been used, as well as a domain-oriented word selection, which have been extracted from the dataset. In this last approach, words were filtered by its frequency of appearance considering the document annotation, being the cutoff frequency an adjustable parameter.

3.2.5 Linguistic Features

The last set of features used are related to linguistic aspects. The proposed system considers the number of sentences, length from the tweet, POS stats, as well as some Twitter-related features such as the count of hashtags, URLs, mentions, all caps words, emojis, and exclamations.

3.3 Classification

Finally, the furthest step in the data processing pipeline makes use of a machine learning classifier. There are many options among machine learning models that can be used. In this project, we have evaluated the performance of three different types of algorithms: Logistic Regression, Support Vector Machines (SVM) with linear kernel, and Random Forest.

4 Experiments

This section presents the results obtained by the proposed system in the competition, considering both test and development phase submissions. Firstly, a data exploration has been carried out in order to analyze the data distribution, possible features to feed the classifier, and deficiencies in the data source. The evaluation of the different feature extraction approaches and the hyper-parameter tuning has been done by using a cross-validation grid search. Special attention has been paid in the regularization parameter of the algorithms: "C" parameter in the Logistic Regression and Linear SVM case and "maximum depth" of the trees in the Random Forest case. Finally, the system is trained, and the evaluation metrics are computed. This workflow has been repeated several times from the feature extraction step, changing the set of features in every iteration.

4.1 Sub-task A

The goal of this task is to classify both Spanish and English tweets as hateful or not hateful. Systems are evaluated using standard evaluation metrics, including accuracy, precision, recall, and F1-score, but predictions are ranked by F1-score metric alone.

Team	Accuracy	Precision	Recall	F-score
English				
Best	0.506	0.65	0.566	0.457
SVM baseline	0.492	0.595	0.549	0.451
GSI-UPM	0.483	0.643	0.549	0.42
MFC baseline	0.579	0.289	0.5	0.367
Spanish				
Best	0.731	0.734	0.741	0.73
GSI-UPM	0.728	0.726	0.733	0.725
SVM baseline	0.705	0.701	0.707	0.701
MFC baseline	0.58	0.294	0.5	0.37

Table 1: Official test set results for Task A

Feature combination	Accuracy	Precision	Recall	F-score
English				
Official submission combination	0.777	0.774	0.780	0.775
Lexical, similarity, embeddings, and n-grams (1)	0.757	0.754	0.758	0.754
Bigrams, trigrams, similarity, and embeddings (2)	0.75	0.747	0.752	0.748
Embeddings, similarity, twitter stats, and LDA (3)	0.736	0.731	0.733	0.732
Spanish				
Official submission combination	0.856	0.856	0.852	0.853
Lexical, similarity, embeddings, and n-grams (1)	0.812	0.811	0.807	0.808
Bigrams, trigrams, similarity, and embeddings (2)	0.796	0.794	0.791	0.792
Embeddings, similarity, Twitter stats, and LDA (3)	0.784	0.781	0.782	0.782

Table 2: Development set results for Task A

Task A data was partitioned into train, development, and test sets. Train and development sets were used to obtain the best feature combination by training over the train set and testing over the development one. Finally, for the final submission, the predictions for the test set were obtained with a system trained over both train and development sets.

Test results, which represent the official submission, as well as development phase results are presented in Tables 1 and 2 respectively. Task organizers included two baselines (Basile et al., 2019) in the competition, a linear SVM based on a TF-IDF representation and a trivial model that assigns the most frequent label from the training set to all instances in the test set.

The Spanish-oriented system relies on linguistic features (excepting POS), semantic similarity with a domain-oriented lexicon, sentiments (using the sentiment vocabulary weighted by the TF-IDF measure), word embeddings, topic modeling (both LDA and hashtags), and word and character TF-IDF n-grams. These features are filtered according to the ANOVA F-test, selecting the best 3,000. Linear SVM has been the selected machine learning algorithm for classification. On the other hand, the English-oriented system considers the same feature set excluding word embedding representation; the number of selected features has been set at 17,500. In contrast to the previous system, a Logistic Regression model was used to perform the classification.

Team	F-score(HS)	F-score(TR)	F-score(AG)	F-score (Avg)	EMR
English					
MFC baseline (Best)	0.367	0.452	0.445	0.421	0.58
GSI-UPM	0.421	0.686	0.556	0.555	0.312
SVM baseline	0.45	0.697	0.587	0.578	0.308
Spanish					
Best	0.729	0.798	0.737	0.755	0.705
GSI-UPM	0.725	0.79	0.735	0.75	0.624
SVM baseline	0.701	0.781	0.726	0.736	0.605
MFC baseline	0.37	0.424	0.413	0.402	0.588

Table 3: Official Results for Task B

Feature Combination	F-score(HS)	F-score(TR)	F-score(AG)	F-score (Avg)	EMR
English					
Official submission	0.775	0.811	0.723	0.770	0.665
(1)	0.754	0.797	0.712	0.755	0.641
(2)	0.748	0.788	0.699	0.745	0.628
(3)	0.731	0.767	0.687	0.728	0.611
Spanish					
Official submission	0.853	0.876	0.824	0.851	0.78
(1)	0.808	0.839	0.777	0.808	0.732
(2)	0.792	0.843	0.776	0.804	0.718
(3)	0.782	0.836	0.783	0.800	0.714

Table 4: Development Results for Task B

4.2 Sub-task B

The goal of this task is firstly to classify hateful tweets (i.e., tweets identified as hate speech against women or immigrants) as aggressive or not aggressive, and secondly to identify the target harassed as individual or generic (i.e., single person or group). Systems are evaluated by two criteria: partial match and exact match (Basile et al., 2019), but predictions are ranked by exact match metric alone.

For this task, the data has been delivered in the same way than sub-task A, so we emulated the same workflow than before, but in this case, considering solely hateful tweets. In this case, there are different distributions (Basile et al., 2019) along languages and sets, but different labels show a similar layout. This result goes in line with the work presented in (ElSherief et al., 2018), which states that directed hate speech is more informal, angrier, and often explicitly attacks the victim. Regarding the language, Spanish-speaking people tend to be more aggressive and more direct towards specific individuals. Seeing this skewed distribution, we outlined the idea to balance aggressiveness and directed messages by oversampling hateful tweets with not hateful ones, assuming that not hateful tweets are not aggressive nor directed.

As done previously, Tables 3 and 4 present official and development results, respectively. The Spanish-oriented system in this task is identical to that from Task A, but finally selecting 2,500 features. For the English case, in light of aggressiveness and target tweets, a different combination of features have been chosen. In order to detect aggressive tweets, all features except semantic similarity have been used, filtering the 32,500 best. Otherwise, for target messages, the complete set of features (sentiments and subjectivity were included by means of TF-IDF and semantic similarity) are used just considering the 2,500 best. Finally, different models were applied for each label, Logistic Regression for Target label and Linear SVM for the Aggressive one. The same algorithm selection was made in the Spanish case.

4.3 Discussion

In general terms, the results obtained are auspicious: the best submitted system achieved the 5th

position in the Spanish Task A, 0.5% points under the best result obtained in the same task. For the Spanish Task B, the proposed model outperforms the baseline. In contrast to this, results in English Task A are lower than expected, where there was not any team that surpassed the 50% threshold in terms of F-score. As a general trend, test set results are worse than development results, which may indicate that our systems suffer over-fitting, and cannot generalize properly. This observation is enforced by attending to the English Task B, where no system has surpassed the baseline.

Since the data distribution is equal along languages in Task A (Basile et al., 2019), the difference in performance across languages may be due to Spanish speaking people are more explicit when typing any utterance with hate speech goals. As previously mentioned, we have observed that this type of hate speech messages show more aggressiveness. Language characteristics may be involved since the Spanish language has a morphologically-richer nature than English.

The presented results constitute the outcome of exhaustive experimentation of a variety of feature combination tests. In contrast with earlier work, semantic similarity and word embeddings representations do not produce such high performance results when compared to other domains such as sentiment analysis (Araque et al., 2019) and sleep disorder detection (Suarez et al., 2018) tasks. This circumstance suggests that hate speech detection is still an open challenge and more research must be done into the specific characteristics of such an exciting task.

Attending to the Spanish case, sentiment information and character n-grams were features that helped in a meaningful manner, confirming the issues raised in Sect. 3. For the English case, the improvement of the proposed features was incremental. While subjectivity and emojis had a relevant role in the results, this improvement was not as high as in the Spanish case. In light of the complexity of the hate speech domain, it could be argued that attending to word context instead of isolated words could help in the analysis. Indeed, n-grams include this type of information to some extent, but capturing the grammatical dependencies within a sentence (Chen, 2011) or template based strategies (Warner and Hirschberg, 2012) could enhance the performance.

5 Conclusions

This paper described the GSI-UPM hate speech detection system presented to participate in SemEval-2019 Task 5, which revolves around analyzing text messages from Twitter. In order to tackle this, a machine learning based approach has been developed. The different features that feed this system have been thoroughly evaluated, considering its suitability in the field of hate speech detection. It has been seen that both novel and traditional approaches do not yield so promising when used separately. Nevertheless, properly combining several types of features, as well as with content analysis features (e.g., sentiments and subjectivity) can improve the system to the point of reaching a reasonably good performance.

Concerning the achieved goals, the highest ranking was 5th place on the Spanish sub-task A, being 0.5% apart from the best performing system. This is, undoubtedly, a promising result that highlights the capacity of the proposed method to obtain nearly state-of-the-art performance in this task. When comparing with the same sub-task in the English case, in which we scored lower, it is necessary to study further the applicability of the system to different languages.

As future work, several lines of work could be addressed. Firstly, we plan to implement deep learning architectures which have shown to obtain better results in previous research (Zhang and Luo, 2018; Zhang et al., 2018). In addition, in order to afford imbalanced distributions, data augmentation (Hemker, 2018) techniques could be explored. Also, context-aware approaches could represent an improvement (Dinakar et al., 2012), since having general knowledge of hate speech (e.g., anti-LGBT or racism) may boost the performance of learning systems.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under the R&D project SEMOLA (TEC2015-68284-R) and the European Union with Trivalent (H2020 Action Grant No. 740934, SEC-06-FCT-2016).

References

Oscar Araque, Ignacio Corcuera-Platas, J. Fernando Snchez-Rada, and Carlos A. Iglesias. 2017. Enhanc-

ing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236 – 246.

Oscar Araque, Ganggao Zhu, and Carlos A. Iglesias. 2019. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 165:346 – 359.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 759–760, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. Nltk: The natural language toolkit. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, ACLdemo '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Ying Chen. 2011. Detecting offensive language in social medias for protection of adolescent online safety.

Keith Cortis and Siegfried Handschuh. 2015. Analysis of cyberbullying tweets in trending world events. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, i-KNOW '15, pages 7:1–7:8, New York, NY, USA. ACM.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst.*, 2(3):18:1–18:30.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth M. Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. *CoRR*, abs/1804.04257.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Konstantin Hemker. 2018. Data augmentation and deep learning for hate speech detection. Master's thesis, Imperial College London.

Michael Herz and Peter Molnar. 2012. *The content and context of hate speech*. Cambridge University Press.

Sarah Hewitt, T. Tiropanis, and C. Bokhove. 2016. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 333–335, New York, NY, USA. ACM.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351, New York, NY, USA. ACM.

Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. 2014. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*.

C.D. Manning, C.D. Manning, H. Schütze, and H.A. SCHUTZE. 1999. *Foundations of Statistical Natural Language Processing*. Mit Press. MIT Press.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

M.F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *ICWSM*, pages 687–690.

D. Suarez, O. Araque, and C. A. Iglesias. 2018. How well do spaniards sleep? analysis of sleep disorders based on twitter mining. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 11–18.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Shuhan Yuan, Xintao Wu, and Yang Xiang. 2016. A two phase deep learning model for identifying discrimination from tweets. In *EDBT*, pages 696–697.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *CoRR*, abs/1803.03662.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web*, pages 745–760, Cham. Springer International Publishing.