MineriaUNAM at SemEval-2019 Task 5: Detecting Hate Speech in Twitter using Multiple Features in a Combinatorial Framework

Luis Enrique Argota Vega
Posgrado en Ciencia e Ingeniería
de la Computación
Universidad Nacional Autónoma de México
Ciudad de México, México
luiso91@comunidad.unam.mx

Jorge Reyes-Magaña
Facultad de Matemáticas
Universidad Autónoma de Yucatán
Mérida, Yucatán, México
Universidad Nacional Autónoma de México
Ciudad de México, México
jorge.reyes@correo.uady.mx

Helena Gómez-Adorno
Instituto de Investigaciones
en Matemáticas Aplicadas
y en Sistemas
Universidad Nacional Autónoma de México
Ciudad de México, México
helena.gomez@iimas.unam.mx

Gemma Bel-Enguix
Grupo de Ingeniería Lingüística
Instituto de Ingeniería
Universidad Nacional Autónoma de México
Ciudad de México, México
gbele@iingen.unam.mx

Abstract

This paper presents our approach to the Task 5 of Semeval-2019, which aims at detecting hate speech against immigrants and women in Twitter. The task consists of two subtasks, in Spanish and English: (A) detection of hate speech and (B) classification of hateful tweets as aggressive or not, and identification of the target harassed as individual or group. We used linguistically motivated features and several types of n-grams (words, characters, functional words, punctuation symbols, POS, among others). For task A, we trained a Support Vector Machine using a combinatorial framework, whereas for task B we followed a multi-labeled approach using the Random Forest classifier. Our approach achieved the highest F1-score in sub-task A for the Spanish language.

1 Introduction

Hate speech is defined as any communication that disparages a person or a group based on some characteristics. Given the enormous amount of content generated by users on the web, and in particular in social networks, the problem of detecting hate speech is becoming fundamental. Early detection of this kind of language can help to limit its dissemination over the web and to fight against misogyny and xenophobia.

The goal of the task (Basile et al., 2019) is to detect hate speech on Twitter in a multilingual perspective, for Spanish and English. The task is divided into two related subtasks for each of the languages: (task A) detection of hate speech, and (task B) identifying whether the objective of hatred is a person or group of people. In addition, this second task questions if the author of the message pretends to be aggressive, harmful or even incites violence, in several aspects.

From a machine learning perspective, the task can be seen as a binary classification problem. In order to solve the tasks, we evaluated several machine learning algorithms: Support Vector Machines, Logistic Regression, Multinomial Naive Bayes, Decision Trees and, Random Forest.

For text representation, we extracted linguistically motivated patterns and several types of n-grams (characters, words, syntactic and, aggressive words, among others). The pre-processing steps and the experiments carried out to solve this task are explained in the following sections.

2 Related work

In recent years, the automatic detection of aggressive behavior in social media is gaining a lot of attention. This is consistent with political and social concern about hatred and harassment through these media. Several evaluation campaigns have been recently organized related to hate speech detection such as the hate speech identification task at Evalita (Bosco et al., 2018), the aggressiveness

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 447–452
Minneapolis, Minnesota, USA, June 6–7, 2019.

detection task at IberEval (Álvarez-Carmona et al., 2018), and the misogyny identification task (Anzovino et al., 2018) at Evalita (Fersini et al., 2018), among many others.

Our work is based on previous work on aggressive detection of tweets in Mexican Spanish (Gómez-Adorno et al., 2018), which was presented in the MEX-A3T 2018 Workshop (Álvarez-Carmona et al., 2018). It follows a classical machine learning approach, a logistic regression algorithm is trained on linguistically motivated characteristics and various types of n-grams (characters, words, syntactic and aggressive words). Furthermore, an oversampling technique (SMOTE) is used to overcome the problem of unbalanced data, which allowed them to achieve better results in the training corpus, but did not generalize well in the test corpus.

When concerning to hate speech detection related methodologies, Djuric et al. (2015) presented a list of criteria based on the critical race theory to identify racist and sexist slander, whereas Chatzakou et al. (2017) implemented a solid methodology for the extraction of text, user and attributes based on a social media network.

Djuric et al. (2015) used the generated list to annotate a publicly available corpus of more than 16k tweets. They analyzed the impact of various extra-linguistic features along with character n-grams for the detection of hate speech. In turn, they elaborated a dictionary based on the most indicative words in their data.

Chatzakou et al. (2017) studied the properties of bullies and aggressors, and the characteristics that distinguish them from normal users. They found that stalkers post with less frequency, participate in fewer online communities and are less popular than users with standard models of behaviour. Their research shows that machine learning classification algorithms can accurately detect users who exhibit bullying and aggressive behavior, with more than 90% of accuracy.

3 Corpus

For the development phase of the competition, the organizers (Basile et al., 2019) distributed the tweets in four files. Statistics referring the corpus for Spanish and English are presented in Table 1.

For the competition evaluation phase, we have put together the training and test corpus of the competition development phase, in table 2 we present details of the resulting corpus.The test corpus of this phase consists of 1,600 tweets for Spanish and 3,000 for English.

4 Methodology

This section shows in detail how tweets are processed for further classification. It is very important the text to be processed in an appropriate format, so that its manipulation can be done in a simpler and less complex way and an optimal precision can be obtained for the automated methods. Additionally, there are several methods for increasing the characteristics of the system, in order to feed the classifier and have more elements when it comes to analyzing your data.

4.1 Pre-processing

Several researchers show that pre-processing is useful for several natural language processing (NLP) tasks (Montes-y-Gómez, 2001; Justicia de la Torre, 2017), especially when the corpus is made of social network data (Pinto et al., 2012; Gómez-Adorno et al., 2016b). Before the extraction of features, the following pre-processing steps are applied in order to improve the representation of n-grams and to reduce the errors of part-of-speech (POS) labeling:

1. The type of single quotes in the English tweets was standardized. This step was prior to the substitution of abbreviations, thus allowing a correct replacement of the equivalent text.

2. All tweets were standardized to lowercase, which avoids having multiple copies of the same words.

3. The mentions to users (@user) were removed.

4. Url's were removed.

5. Emojis were removed.

6. Following the methodology of Gómez-Adorno et al. (2016a), the abbreviations, contractions and, slangs were replaced by the equivalent text for both Spanish and English. It is important to mention that the vocabulary used in these lexical resources is based on social networks.

		Hateful	Individual target	Aggressive	Total
Spanish	Training	1,838 (41.12%)	1,117 (24.99%)	1,485 (33.22%)	4,469
	Testing	222(44.4%)	137 (27.4 %)	176(35.19%)	500
English	Training	3,783 (42.03%)	1,341 (14.89%)	1,559 (17.32%)	9,000
	Testing	427(42.7%)	219 (21.9 %)	204(20.40%)	1,000

Table 1: Corpus statistics for Task A and B in the development phase

	Hateful	Individual target	Aggressive	Total
Training Spanish	2,060 (41.45%)	1,254 (25.23%)	1,661 (33.42%)	4,969
Training English	4,210 (42.10%)	1,560 (15.60%)	1,763 (17.63%)	10,000

Table 2: Corpus statistics for Task A and B in the evaluation phase

7. Function words (or stopwords) were removed.

8. We replaced the figures that appear in tweet by a single digit (0), since the numbers do not contain semantic information that could be relevant for the task.

9. For hashtags, we had the following criteria: if there was a word detected as hateful/aggressive in the text of the hashtag, the complete hashtag was replaced by that word. If no sign of aggressiveness/hatred was found, then the hashtag was removed.

10. Certain rare and special characters were detected in the tweets and they were replaced by a blank space.

11. In the tweets in English, words such as "&amp" or the character "&" were detected, which represented a conjunction, in which case it was replaced by the word "and".

12. We deleted punctuation, since it does not add any additional information when processing text data. Therefore, eliminating all cases helps to reduce the size of the training and test data.

13. The sequences of several blank spaces, tabs and line breaks were standardized to a single blank space.

4.2 Features

We took into account several features for the representation of tweets:

- **Character n-grams.** Are capable of detecting the morphological composition of a word (Kulmizev et al., 2017). For natural language processing tasks, where many words are likely to be poorly written, the n-grams of characters are especially powerful (Sanchez-Perez et al., 2017) to detect patterns in such spelling mistakes (Kulmizev et al., 2017). For this approach, a variation of n from 3 to 5 is included.

- **Word n-grams.** Capture the identity of a word and its possible neighbors (Kulmizev et al., 2017). In the experiments, the combination of the n-grams with n varying from 1 to 4 helps to improve the results.

- **POS tags n-grams.** Are sequences of continuous part-of-speech (POS) tags. They capture syntactic information and are useful, for example, to identify the user's intentions in tweets (Gómez-Adorno et al., 2018). We have experimented with various combinations of POS n-grams in the data, finding that a range of 2 to 4 provides the best results in the development set.

- **Aggressive word n-grams.** For this work, we gathered a lexicon of aggressive words containing those words obtained by (Gómez-Adorno et al., 2018) and other words we extracted from the training corpus. We built bigrams and trigrams only with the words of this lexicon.

- **Skipgrams.** We capture groups of 2 words with skips of 2 to 4 words.

- **Function words n-grams.** The frequency of this words is one of the best characteristics to detect hate speech and aggressiveness. Prior

to the pre-processing of the corpus, we built function words n-grams from 2 to 4 tokens for both languages. We used the stopwords list from NLTK.

- **N-grams of punctuation symbols.** With this feature we approached the coherence and cohesion to the written text. It helps to detect certain patterns in the analysis of hatred and aggressiveness. Prior to the corpus preprocessing, we built n-grams of 2 to 4 punctuation symbols.

- **Language patterns.** We performed a linguistic analysis of the entire training corpus to detect language patterns that can help to distinguish if the tweets are directed to a person or group of people. We considered two types of patterns: morphological structures, and recurrent lexical patterns.

  - Gómez-Adorno et al. (2018) established that certain morphological combinations can help the classification of tweets. Taking into account this technique, the following combinations were detected: verb + adjective, adjective + verb, noun + adjective, adjective + noun and pronoun + verb.
  - Lexical patterns formed with the series of aggressive words detected in the tweets.

4.3 Classifier

For task A, we used a combinatorial framework (μTC) developed by Tellez et al. (2018). The framework approaches any text classification problem as a combinatorial optimization problem; where there is a search space containing all possible combinations of different text transformations (tokenizers) and weighting schemes with their respective parameters, and, on this search space, a meta-heuristic is used to search for a configuration that produces a highly effective text classifier. Considering all the combinations established in the implementation[1] of (μTC), we added the features described in Section 4.2 and the preprocessing techniques in Section 4.1. Once, the best feature space, we trained an SVM with linear kernel.

For task B, we used the sklearn.multiclass[2] module that implements meta-estimators to solve multi-class and multi-label classification problems, decomposing these problems in binary classification problems. In this sense, the multi-label classification assigns a set of target labels to each sample. In particular, we used the Random Forest algorithm, which showed a better performance than the other algorithms of machine learning that we examined. We performed 10-fold cross-validation experiments to select the best features, weighting scheme and, frequency threshold. The final configuration of the system implements a binary weighting scheme, and considers only those characteristics that occur at least 10 times throughout the corpus and that occur in at least 50 documents in the corpus.

5 Results

The performance measure used for task A is the F1 score. Table 3 and Table 4 show the results using the SVM algorithm for both cases, the development phase and the official results of the final evaluation phase. We obtained the best overal F1 score for task A in Spanish, however, for task A in English the results were much lower.

Position	Team	Dev	Eva
1	Atalaya	-	0.73
1	**mineriaUNAM**	0.80	0.73
3	MITRE	-	0.729

Table 3: Results of task A in Spanish of the development phase (Dev) and the official results of the final evaluation phase (Eva).

Position	Team	Dev	Eva
1	Fermi	-	0.651
2	Panaetius	-	0.571
3	YNU_DYX	-	0.546
...	...	...	...
60	**mineriaUNAM**	0.72	0.384

Table 4: Results of task A in English of the development phase (Dev) and the official results of the final evaluation phase (Eva)

The performance measure used for task B is the Exact Match (EMR). Table 5 and Table 6 show the results using the Random Forest algorithm in the

[1] https://github.com/INGEOTEC/microtc

[2] https://scikit-learn.org/stable/

development phase, as well as the official results of the final evaluation phase. For this sub-task we achieved better results in the English language than in Spanish.

Position	Team	Dev	Eva
1	CIC-2	-	0.705
2	CIC-1	-	0.675
3	MITRE	-	0.671
...	...	...	...
16	**mineriaUNAM**	0.73	0.596

Table 5: Results of task B in Spanish of the development phase (Dev) and the official results of the final evaluation phase (Eva).

Position	Team	Dev	Eva
1	MFC baseline	-	0.58
2	LT3	-	0.57
3	CIC-1	-	0.568
...	...	...	...
11	**mineriaUNAM**	0.54	0.368

Table 6: Results of task B in English of the development phase (Dev) and the official results of the final evaluation phase (Eva).

6 Conclusions

We presented an approach for the detection of hate speech, aggressiveness and harassed objective as individual or group in Twitter in Spanish and English.

We implemented a Support Vector Classification algorithm since it presented a better prediction with respect to other methods. In the case of multi-label classification, the Random Forest algorithm was used. Both algorithms were trained on a combination of linguistic patterns features, a lexicon of aggressive words and different types of n-grams (characters, words, POS tags, aggressive words, word jumps, function words, and punctuation symbols).

The results obtained in Task A when using only the μTC original implementation were improved by the addition of extra features such as the aggressive words n-grams, the punctuation symbols n-grams, and the function words n-grams. Remember that the μTC framework is composed of several easy-to-implement text transformations which allowed us to obtain a high-performance classification model. The main advantage of this technique is that it automatically adjusts the parameters of the model.

Acknowledgments

This paper has been supported by project PAPIIT IA401219 from the Universidad Nacional Autónoma de México.

References

Miguel Á Álvarez-Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, Verónica Reyes-Meza, and Antonio Rico-Sulayes. 2018. Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *Notebook Papers of 3rd. SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September*.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

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.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22. ACM.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA18), Turin, Italy.*

Helena Gómez-Adorno, Gemma Bel-Enguix, Gerardo Sierra, Octavio Sánchez, and Daniela Quezada. 2018. A machine learning approach for detecting aggressive tweets in spanish. In *In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings*.

Helena Gómez-Adorno, Ilia Markov, Grigori Sidorov, Juan Pablo Posadas-Durán, and Carolina Fócil Arias. 2016a. Compilación de un lexicón de redes sociales para la identificación de perfiles de autor. *Research in Computing Science*, 115:19–27.

Helena Gómez-Adorno, Ilia Markov, Grigori Sidorov, Juan-Pablo Posadas-Durán, Miguel A Sanchez-Perez, and Liliana Chanona-Hernandez. 2016b. Improving feature representation based on a neural network for author profiling in social media texts. *Computational intelligence and neuroscience*, 2016:2.

Artur Kulmizev, Bo Blankers, Johannes Bjerva, Malvina Nissim, Gertjan van Noord, Barbara Plank, and Martijn Wieling. 2017. The power of character n-grams in native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 382–389.

Manuel Montes-y-Gómez. 2001. Minería de texto: Un nuevo reto computacional.

David Pinto, Darnes Vilarino, Yuridiana Alemán, Helena Gómez, and Nahun Loya. 2012. The soundex phonetic algorithm revisited for sms-based information retrieval. In *II Spanish Conference on Information Retrieval CERI*.

Miguel A Sanchez-Perez, Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov. 2017. Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 145–151. Springer.

Eric S. Tellez, Daniela Moctezuma, Sabino Miranda-Jiménez, and Mario Graff. 2018. An automated text categorization framework based on hyperparameter optimization. *Knowledge-Based Systems*, 149:110–123.

María del Consuelo Justicia de la Torre. 2017. Nuevas técnicas de minería de textos: Aplicaciones.