CiTIUS-COLE at SemEval-2019 Task 5: Combining Linguistic Features to Identify Hate Speech Against Immigrants and Women on Multilingual Tweets

Sattam Almatarneh
(CiTIUS)
Universidade de Santiago
de Compostela, Spain
University of Vigo, Spain
sattam.almatarneh@usc.es

Pablo Gamallo
(CiTIUS)
Universidade de Santiago
de Compostela, Spain
pablo.gamallo@usc.es

Francisco J. Ribadas Pena
Department of Computer Science
University of Vigo, Spain
ribadas@uvigo.es

Abstract

This article describes the strategy submitted by the CiTIUS-COLE team to SemEval 2019 Task 5, a task which consists of binary classification where the system predicts whether a tweet in English or in Spanish is hateful against women or immigrants or not. The proposed strategy relies on combining linguistic features to improve the classifier's performance. More precisely, the method combines textual and lexical features, embedding words with the bag of words in Term Frequency-Inverse Document Frequency (TF-IDF) representation. The system performance reaches about 81% F1 when it is applied to the training dataset, but its F1 drops to 36% on the official test dataset for the English and 64% for the Spanish language concerning the hate speech class.

1 Introduction

Hate speech is usually defined as any communication that derogates a person or a group based on some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or another characteristic (Schmidt and Wiegand, 2017). The spread of the Internet and the increasing use of social networks has led people to have an increased willingness to express their opinions online. Despite the great benefits of using the Internet and particularly the social networks, the risk is that people are more likely to adopt aggressive behavior because of the anonymity provided by these environments. This contributes to the propagation of hate speech as well. Since this type of tendentious communication can be extremely harmful to society, governments and social network platforms can benefit from detection and prevention tools. The scientific study of hate speech, from a computer science point of view, is recent, and the number of studies in the field is low (Fortuna and Nunes, 2018). The goal of SemEval-2019 Task 5 as described in Basile et al. (2019) is Hate Speech detection in Twitter focused on two specific targets, women and immigrants. The task is organized in two related sub-tasks for each language (English and Spanish):

- TASK A - Hate Speech Detection against Immigrants and Women
- TASK B - Aggressive behavior and Target Classification.

In this article, we describe our proposed system for task A only. Our approach is mainly based on the generation of corpus-based dictionaries containing hate speech words which are used in addition to other linguistic features to improve the efficiency in detecting hate speech in both English and Spanish languages.

This paper is organized as follows. The method is described in Section 2. Experiments, results, and a discussion on them are presented in Section 3. Finally, conclusions are addressed in Section 4.

2 Method

We deal with the task by automatic classifiers composed of training data in a supervised strategy. The characteristics of tweets are encoded as features in vector representation. These vectors and the corresponding labels feed the classifiers.

2.1 Features

Linguistic features are the most important and influential factor in increasing the efficiency of classifiers for any task of text mining. Many studies examined the impact of these features in many tasks such as polarity classification (Almatarneh and Gamallo, 2018b, 2019). In this study, we included a number of linguistic features for the task of determining hate speech in tweets. The main

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 387–390
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

linguistic features we will use and analyze are the following: N-grams, word embeddings, and lexical features.

2.1.1 TF-IDF features

We model texts by n-grams based on the occurrence of unigrams of words that occur in documents. The unigrams are very valuable elements to find very relevant expressions in the domain of interest. All terms are assigned a weight by TF-IDF which is computed in Equation 1:

$$tf/idf_{t,d} = (1 + log(tf_{t,d})) \times log(\frac{N}{df_t}), \quad (1)$$

where $tf_{t,d}$ is the term frequency of term t in document d. N stands for the the number of documents in the collection and, df_t represents the number of documents in the collection containing t. To transform the tweets into a matrix of TF-IDF features, we used *sklearn* feature extraction Python library.[1]

2.1.2 Doc2Vec

To represent the tweets, we make use of the *Doc2Vec* algorithm described in Le and Mikolov (2014). This neural-based model is efficient when you have to account for high-dimensional and sparse data (Le and Mikolov, 2014; Dai et al., 2015). Doc2vec learns corpus features using an unsupervised strategy and provides a fixed-length feature vector as output. The output is then fed into a machine learning classifier. We used a freely available implementation of the Doc2Vec algorithm included in gensim, [2] which is a free Python library. The implementation of the Doc2Vec algorithm requires the number of features to be returned (length of the vector). Thus, we performed a grid search over the fixed vector length 100 (Collobert et al., 2011; Mikolov et al., 2013a,b).

2.1.3 Lexical features

Lexical features consist of specific words identified as belonging to the class of hate speech. For instance, as word *bitch* can be associated with hate speech, it will be added to a specific dictionary containing words associated with hate speech. In addition, a weight is assigned to each word. The higher the weight the more intense the hate value of the word. We automatically built several weighted dictionaries from the annotated corpus:

- Dictionary of lexical words 295 English words and 262 Spanish words.
- Dictionary of hashtags: 1090 English hashtags and 201 Spanish hashtags.
- Dictionary of address references: 1661 English references and 1263 Spanish references.

We just considered words belonging to lexical categories, hence, only nouns, verbs, adjectives, and adverbs were selected. PoS tagging for English and Spanish was carried out with the multilingual toolkit LinguaKit (Gamallo et al., 2018).

The method to build the hate speech dictionaries is somehow inspired by that reported in Almatarneh and Gamallo (2018a, 2017) for very negative opinions. The hate speech score of a word, noted HS, is computed as follows:

$$HS(w) = \frac{freq_{total}(w)}{freq_{hs}(w)} \quad (2)$$

where $freq_{total}(w)$ is the number of occurrences of word w in the whole corpus, and $freq_{hs}(w)$ stands for the number of occurrences of the same word in the segments (tweets) annotated as hate speech. In addition to the hate speech score HS, it is also required to compute a threshold above which the word is considered hate speech. So, we compute the difference between the use of a word as hate speech and as not:

$$DIFF(w) = freq_{hs}(w) - freq_{-hs}(w) \quad (3)$$

where $freq_{-hs}(w)$ stands for the occurrences of w in segments that are not hate speech. To insert a word in the dictionary, the value of $DIFF(w)$ must be higher than a experimentally set threshold. In our experiments, this value was 5. So, in our dictionaries, we only selected those words (hashtags or references) with $DIFF$ values higher than 5. Finally, words were ranked by their HS score giving rise to weighted and ranked lexicons.

3 Experiments

The main datasets that were used for training and testing our model are described in Basile et al. (2019). This article describes the SemEval-2019 Shared Task 5 aimed at Hate Speech detection in Twitter.

[1] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html#sklearn.feature_extraction.text.TfidfVectorizer

[2] https://radimrehurek.com/gensim/

Table 1: Performance on the training dataset with different feature configurations of TF-IDF, Doc2Vec and lexicons for English language.

Featuers	Hate			Not			Avg F1	Accuracy
	Prec.	Rec.	F1	Prec.	Rec.	F1		
TF-IDF	0.72	0.68	0.70	0.79	0.82	0.80	0.76	0.76
Doc2Vec	0.68	0.52	0.59	0.71	0.83	0.77	0.69	0.70
Lexicon	0.69	0.46	0.55	0.7	0.86	0.77	0.68	0.69
Doc2Vec + Lexicon	0.71	0.57	0.63	0.74	0.84	0.79	0.72	0.73
All	0.78	0.74	**0.76**	0.83	0.86	**0.84**	**0.81**	**0.81**

Table 2: Performance on the training dataset with different feature configurations of TF-IDF, Doc2Vec and lexicons for Spanish.

Featuers	Hate			Not			Avg F1	Accuracy
	Prec.	Rec.	F1	Prec.	Rec.	F1		
TF-IDF	0.75	0.7	0.72	0.78	0.83	0.80	0.77	0.77
Doc2Vec	0.57	0.11	0.18	0.58	0.94	0.72	0.49	0.58
Lexicon	0.82	0.66	0.73	0.78	0.89	0.83	0.79	0.79
Doc2Vec + Lexicon	0.82	0.68	0.74	0.78	0.89	0.83	0.79	0.80
TF-IDF + Lexicon	0.81	0.76	**0.78**	0.83	0.86	**0.85**	**0.82**	**0.82**

3.1 Development and training

As we considered that the size of the training collection provided by the organizers was not large enough, we made use of another available training data for the same task to build our lexicons.[3] The algorithm to build the lexicons has been described above in Subsection 2.1.3.

As far as the classification strategy is concerned, we decided to use *sklearn.svm.LinearSVC* for learning the classifiers.[4] Suport Vector Machine (SVM) proved to be the best strategy for detecting extreme opinions in previous work (Almatarneh and Gamallo, 2019)

The training dataset provided by the organizers of the shared task was used as a development corpus so as to learn the best feature configuration using 10-fold cross-validation.

Tables 1 and 2 shows the result of the experiments on the training corpus. In these tables, we depict the performance of all tested features in both English and Spanish Languages. The combination of all features (TF-IDF, Doc2Vec, and lexicons) gives the best performance for English. However, in Spanish the use of Doc2Vec made it lower the performance as the best F1 was achieved by just combining TF-IDF with lexical features.

3.2 Test

Taking into account the results shown in tables 1 and 2, we submitted two different model configurations for English and Spanish testing. More precisely, for English TASK A we used the combination of all features, whereas the Spanish model in TASK A was built by only combining lexical and TF-IDF features.

Unlike the experiments on the training dataset, our approach showed bad performance on the test dataset as Table 3 shows.

The poor scores in the English dataset are due to the strange behavior of our approach with the non-hate speech class of speech class. Recall on this class was merely 0.07 while on the target class reached 0.97.

4 Conclusions and Future Work

The approach we developed for the task of hate speech detection in English and Spanish is mainly based on the generation of lexicons containing hate speech words. Lexicons are used in addition to other linguistic features (TF-IDF and Doc2Vec) to improve the efficiency of a SVM classifier.

[3] https://github.com/ZeerakW/hatespeech/blob/master/NAACL_SRW_2016.csv

[4] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

Table 3: Performance of our approach on the test dataset for English and Spanish languages

Featuers	Hate			Not			Avg F1	Accuracy
	Prec.	Rec.	F1	Prec.	Rec.	F1		
TASK A English	0.43	0.97	0.60	0.77	0.07	0.13	0.36	0.45
TASK A Spanish	0.59	0.54	0.56	0.70	0.74	0.72	0.64	0.66

Even if we obtained acceptable results in the development phase using the training corpus (more than 0.80 F1 score), the results achieved in the test phase were disappointing, especially for the English language.

In order to discover the problems underlying our overfitted model, a deep error analysis will be performed. Once released the dataset test, we will be able to analyze the contribution of each of the features used so that we can check if it was one of the lexicons that caused the low performance of our system.

In future work, our objective is to improve the basic method to build hate speech lexicons (and related topics) from annotated corpora in order to use them in both supervised and unsupervised strategies.

Acknowledgments

This work has received financial support from project TelePares (MINECO, ref:FFI2014-51978-C2-1-R), and the Consellera de Cultura, Educacin e Ordenacin Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

References

Sattam Almatarneh and Pablo Gamallo. 2017. Automatic construction of domain-specific sentiment lexicons for polarity classification. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 175–182. Springer.

Sattam Almatarneh and Pablo Gamallo. 2018a. A lexicon based method to search for extreme opinions. *PloS one*, 13(5):e0197816.

Sattam Almatarneh and Pablo Gamallo. 2018b. Linguistic features to identify extreme opinions: An empirical study. In *Intelligent Data Engineering and Automated Learning – IDEAL 2018*, pages 215–223, Cham. Springer International Publishing.

Sattam Almatarneh and Pablo Gamallo. 2019. Comparing supervised machine learning strategies and linguistic features to search for very negative opinions. *Information*, 10(1).

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, location = Minneapolis, Minnesota.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

Andrew M Dai, Christopher Olah, and Quoc V Le. 2015. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

P. Gamallo, M. Garcia, C. Pieiro, R. Martinez-Castao, and J. C. Pichel. 2018. Linguakit: A big data-based multilingual tool for linguistic analysis and information extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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.