SINAI at SemEval-2019 Task 5: Ensemble learning to detect hate speech against inmigrants and women in English and Spanish tweets

Flor Miriam Plaza-del-Arco, M. Dolores Molina-González,
M. Teresa Martín-Valdivia, L. Alfonso Ureña-López

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{fmplaza, mdmolina, maite, laurena}@ujaen.es

Abstract

Misogyny and xenophobia are some of the most important social problems. With the increase in the use of social media, this feeling of hatred towards women and immigrants can be more easily expressed, therefore it can cause harmful effects on social media users. For this reason, it is important to develop systems capable of detecting hateful comments automatically. In this paper, we describe our system to analyze the hate speech in English and Spanish tweets against Immigrants and Women as part of our participation in SemEval-2019 Task 5: hatEval. Our main contribution is the integration of three individual algorithms of prediction in a model based on *Vote* ensemble classifier.

1 Introduction

With the growing prominence of social media like Twitter or Facebook, more and more users are publishing content and sharing their opinions with others. Unfortunately, the content often contains hate speech language that can have damaging effects on social media users. This fact concerns to social media platforms like Facebook since according to an EU's report, it removes 82 percent of illegal hate speech on the platform, up from 28 percent in 2016[1].

Normally, hate speech can be aimed at a person or a group base on some characteristic such as race, sexuality, color, ethnicity, physical appearance, religion, among others (Erjavec and Kovačič, 2012). Currently, two of the targets most affected by these types of offensive comments are immigrants and women (Waseem and Hovy, 2016). In particular, when the hate speech is gender-oriented, and it specifically targets women, we refer to it as misogyny (Manne, 2017) and when the hate speech is against immigrants, we refer to it as xenophobia (Sanguinetti et al., 2018).

Recently, a growing number of researchers have started to focus on studying the task of automatic detection of hateful language online (Fortuna and Nunes, 2018; Fersini et al., 2018b), moreover, some academic events and shared tasks have taken place focusing on this issue (Fersini et al., 2018a). It is consider as a difficult task for social media platforms. For example, popular social media such as Twitter, Instagram or Facebook are not able to automatically solve this problem and depend on their community to report hateful speech content.

The severe consequences of this problem, combined with the large amount of data that users publish daily on the Web, requires the development of algorithms capable of automatically detecting inappropriate online remarks.

In this paper, we describe our participation in SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (hatEval) (Basile et al., 2019). In particular, we participate in task A in English and Spanish. It is a binary classification task and the objective is predict whether a tweet with a given target (women or immigrants) is hateful or not hateful.

The rest of the paper is structured as follows. In Section 2 we explain the data used in our methods. Section 3 presents the details of the proposed systems. In Section 4, we discuss the analysis and evaluation results for our system. We conclude in Section 6 with remarks and future work.

2 Data

To run our experiments, we used the Spanish and English datasets provided by the organizers in SemEval19 Task 5 : HatEval (Basile et al., 2019). The datasets contain tweets with several fields. Each tweet is composed for an identifier (id), the

[1] https://cnb.cx/2RGmEwel

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 476–479
Minneapolis, Minnesota, USA, June 6–7, 2019.

text of the tweet (text), the mark of hate speech (HS), being 0 if the text is not hateful and 1 if the text is hateful, the mark of recipient of text (TR), being 1 if the target is a single human and 0 if the target is a group of persons and the last field (AG) is the mark that identifies if the text is aggressive whose value is 1, else 0 in the case opposite. During pre-evaluation period, we trained our models on the train set, and evaluated our different approaches on the dev set. During evaluation period, we trained our models on the train and dev sets, and tested the model on the test set. Table 1 shows the number of tweets used in our experiments for Spanish and English.

Dataset	train	dev	test
Spanish	4,500	500	1,600
English	9,000	1,000	3,000

Table 1: Number of tweets per HatEval dataset

We only take into account the fields text and HS for our experiments because we participate in task A in English and Spanish.

3 System Description

In this section, we describe the systems developed for the Hateval task 5, subtask A in English and Spanish.

3.1 Our classification model

In first place, we preprocessed the corpus of tweets provided by the organizers. We applied the following preprocessing steps: the documents were tokenized using NLTK library [2] and all letters were converted to lower-case. In second place, an important step is converting sentences into feature vectors since it is a focal task of supervised learning based sentiment analysis method. Therefore, our chosen statistic feature for the text classification was the term frequency (TF) taking into account unigrams and bigrams because it provided the best perfomance.

During our experiments, the scikit-learn machine learning in Python library (Pedregosa et al., 2011) was used for benchmarking. Our classification model based on *Vote* ensemble classifier combined three individual algorithms: *Logistic Regression (LR)*, *Decision Tree (DT)* and *Support Vector Machines (SVMs)*. We have tested with other models such as *naive bayes* and *multilayer perceptron* but we have obtained better results with the combination of the three algorithms mentioned above. In Figure 1, it can be seen our model. We train our model with the train and dev set and we evaluated it with the test set. There are many combinations to implement a model when we apply different classifiers with several parameters. Therefore, one of the most important step was to find the best individual classifiers for the problem. After doing several experiments with each classifier independently, we came up with SVMs, LR and DT classifiers. In order to improve the performance of each classifier, we choose the best optimization of the parameters in each of them.

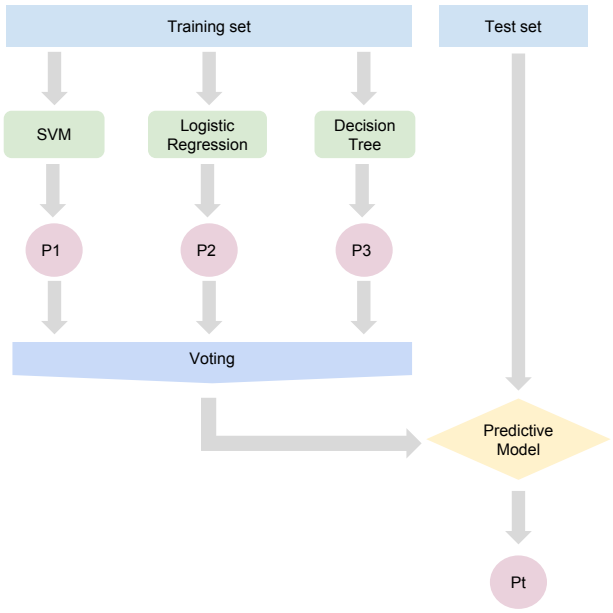

Figure 1: Systems architecture.

3.2 Classifiers

1. *Logistic Regression* is an statistical method for prediction binary classes. It computes the probability of an event occurrence utilizing a logit function. In order to optimize the parameters of LR in our English and Spanish experiments, we used the penalty parameter equal to l1 regularization.

2. *Decision Tree* is a flowchart-like tree structure where an internal node represents features, the branch represents a decision rule, and each leaf node represents the outcome. In order to optimize the parameters of DT in our English and Spanish experiments, we leave

[2] https://www.nltk.org/

User name (r)	Test			
	P	R	F1	Acc
francolq2 (1)	0.734	0.741	0.73	0.731
luiso.vega (2)	0.729	0.736	0.73	0.734
fmplaza (14)	**0.707**	**0.713**	**0.707**	**0.711**
SVC baseline (21)	0.701	0.707	0.701	0.705
DA-LD-Hildesheim (40)	0.493	0.494	0.493	0.511

Table 2: System Results per team in subtask A of hatEval task in Spanish.

User name (ranking)	Test			
	P	R	F1	Acc
saradhix (1)	0.69	0.679	0.651	0.653
amontejo (5)	0.601	0.577	0.519	0.535
SVC baseline (35)	0.595	0.549	0.451	0.492
fmplaza (40)	**0.627**	**0.555**	**0.443**	**0.493**
sabino (71)	0.652	0.521	0.35	0.447

Table 3: System Results per team in subtask A of hatEval task in English.

the default parameters.

3. *Support Vector Machines* is a linear learning technique that finds an optimal hyperplane to separate our two classes (hateful and not hateful speech). Many researchers have reported that this classifier is perhaps the most accurate method for text classification (Moraes et al., 2013) and also is widely used in sentiment analysis (Tsytsarau and Palpanas, 2012). In order to optimize the parameters of SVMs in our English and Spanish experiments, we used the parameter C equal to 0.6 and the kernel used was linear.

4. *Vote* is one of the most straightforward ensemble learning techniques in which performs the decision process by applying several classifiers. Voting classifier combines machine learners by using a majority vote or predicted probabilities for the classification of samples. The predictions made by the submodels can be assigned weights. In our case, the weights are distributed as follows: 2 for LR and SVM and 1 for DT.

4 Experiments and analysis of results

During the pre-evaluation phase we carried out several experiments and the best experiments were taken into account for the evaluation phase. The system has been evaluated using the official competition metrics, including Accuracy (Acc), Precision (P), Recall (R) and F1-score (F1). The metrics have been computed as follows:

$$P = \frac{\text{number of correctly predicted instances}}{\text{number of predicted labels}} \quad (1)$$

$$R = \frac{\text{number of correctly predicted labels}}{\text{number of labels in the gold standard}} \quad (2)$$

$$F1 = \frac{2 * P * R}{P + R} \quad (3)$$

$$Acc = \frac{\text{number of correctly predicted instances}}{\text{total number of instances}} \quad (4)$$

The results of our participation in the subtask A of hatEval task during the evaluation phase can be seen in Table 2 for Spanish and in Table 3 for English.

In relation to Spanish results, it should be noted that we achieve a high position in the ranking outperforming the baseline result. Our position in the ranking is 14th of 41 participating teams. Therefore, we consider that the chosen individual classifiers in the voting system are appropriate to build the metaclassifier.

Therefore, our chosen statistic feature for the text classification was the term frequency (TF) taking into account unigrams and bigrams because it provided the best perfomance. One important feature to consider is the use of bigrams in TF, because during the pre-evaluation phase we noted that our results outperformed when we took into account the bigrams comparing it only to the unigrams.

In relation to English results, using the same system as for Spanish we achieved worse results and we did not outperform the baseline. However, we are ranked 40th out of 71 participating teams.

5 Conclusions

In this paper, we present the system we developed for our participation in SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (hatEval). Specially, we have participated in subtask A in Spanish and English.

Our system was developed focus on Spanish. Therefore, we achieve better results in this language. On the one hand, one of the reasons could be the different employment of misogynistic or xenophobic words in one language with respect to the other (Canós, 2018). For example, the word "puta" in Spanish, can be consider a misogynistic word or in a bigram like "puta madre" can be similar to the word "fantastic". On the other hand, the way to insult women is not the same as the way to insult immigrants. For these reasons, systems make mistakes and should be considered different systems for these targets (immigrants and women).

Another important issue is that the participation in Spanish subtask is lower than the participation in English subtask. For this reason, we will continue developing systems in Spanish since it is one of the most spoken languages in the world and we consider a very challenging task.

Acknowledgments

This work has been partially supported by Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

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.

Jose Sebastián Canós. 2018. Misogyny identification through svm at ibereval 2018.

Karmen Erjavec and Melita Poler 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.

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

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on automatic misogyny identification at ibereval 2018.

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

Kate Manne. 2017. *Down girl: The logic of misogyny*. Oxford University Press.

Rodrigo Moraes, JoãO Francisco Valiati, and Wilson P GaviãO Neto. 2013. Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications*, 40(2):621–633.

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.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Stranisci Marco. 2018. An italian twitter corpus of hate speech against immigrants. In *Language Resources and Evaluation Conference-LREC 2018*, pages 1–8. ELRA.

Mikalai Tsytsarau and Themis Palpanas. 2012. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514.

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.