Ensemble Learning to Detect Aggressiveness in Mexican Spanish Tweets

María Dolores Molina-González, Flor Miriam Plaza-del-Arco, María Teresa Martín-Valdivia, and Luis Alfonso Ureña-López

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

Abstract. Comments published on social media often contain aggressive language that can have damaging effects on users. The severe consequences of this problem, combined with the large amount of data that users daily publish on the Web, require the development of algorithms capable of automatically detecting inappropriate online remarks. In this paper, we present our participation in IberLEF-2019: subtask MEX-A3T: Authorship and aggressiveness analysis in Twitter: case study in Mexican Spanish. Our main contribution is the development of a ensemble learning system to detect aggressiveness in tweets.

Keywords: automatic aggressiveness detection · classifier ensemble · machine learning · social media · text mining

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. This content has the potential to be transmitted quickly, reaching anywhere in the world in few seconds. Unfortunately, the comments often contain aggressiveness language that can have damaging effects on social media users. The hate speech detection includes different issues, such as: misogyny, xenophobia, homophobia, cyberbullying, nastiness and aggressiveness. One of the strategies used to deal with these online hateful behaviors and attitudes in social media is reporting or monitoring this type of content with the main aim of limiting it. However, it is difficult to monitor efficiently and automatic support techniques should be used.

Recently, a growing number of researchers have started to focus on studying the task of automatic detection of hateful language online [6]. Moreover, some national and international workshops and campaigns of evaluation have taken

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). IberLEF 2019, 24 September 2019, Bilbao, Spain.

CEUR-WS.org/Vol-2421/MEX-A3T_paper_1.pdf

place focusing on the research in this issue in various languages, such as the first and second editions of the Workshop on Abusive Language [9], the First Workshop on Trolling, Aggression and Cyberbullying [7], which also included a shared task on aggression identification, the tracks on Automatic Misogyny Identification (AMI) [5] and on authorship and aggressiveness analysis (MEX-A3T) [1] proposed at the 2018 edition of IberEval, the GermEval Shared Task on the Identification of Offensive Language [10], the Automatic Misogyny Identification task at EVALITA 2018 [4], and finally the SemEval shared task on HS detection against immigrants and women (HatEval) [3].

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

In this paper, we describe our participation in IberLEF-2019: subtask MEX-A3T: Authorship and aggressiveness analysis in Twitter: case study in Mexican Spanish [2]. This track proposes to detect the aggressiveness on Mexican Spanish tweets providing texts containing offensive messages that disparage or humiliate specific target.

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 5 with remarks and future work.

2 Data

To run our experiments, we used the Mexican Spanish datasets provided by the organizers in IberLEF-2019 subtask MEX-A3T: Authorship and aggressiveness analysis in Twitter: case study in Mexican Spanish [2]. The dataset description contains two files: one of them contains 7,700 Mexican Spanish tweets of the training set (one tweet per line) and the other one contains the corresponding labels of the 7,700 tweets of the training set (one label per line).The label has two possible classes: 0 means "non-aggressive", 1 means "aggressive". The 7,700 tweets have been processed before releasing. The organizers have changed all user mentions as @USUARIO.

During pre-evaluation period, we trained our models on the train set, and evaluated different approaches with 10-fold cross-validation. During evaluation period, we trained our models on the train and tested the model on the test set. Table 1 shows the number of tweets used in our experiments.

Table 1: Number of tweets per MEX-A3T dataset

Dataset	Non AG	AG	Total
Train	4,973	2,727	7,700
Test	-	-	3,156

3 System Description

In this section, we describe how we addressed the identification of aggressiveness in Twitter, and in particular MEX-A3T organizers proposed a classification task with the aim to distinguish aggressive tweet from the non-aggressive from Mexican Spanish users.

3.1 Our classification model

In first place, we preprocessed the corpus of tweets provided by the organizers. After the tokenization process, we carried out the following steps:

- Lower-case conversion data.
- Normalize URLs, emails, users mentions, percent, money, time, date expressions and phone numbers.
- Unpack hashtags (e.g. *#HechosReales* becomes <hashtag>*hecho reales* <hashtag>).
- Annotate and reduce elongated words (e.g. *agresivooooooo* becomes <elongated>*agresivo*) and repeat words (e.g. !!!! becomes <repeated>!).
- Map emoticons.

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 performance.

During our experiments, the scikit-learn machine learning library in Python [8] was used for benchmarking.

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 classifier for the problem. Table 2 shows the results associated with each evaluated classifier in the training phase.

Table 2: Systems Results of train set

Classifier	Acc	P (1)	P (0)	R (1)	R (0)	F1 (1)	R (avg)	F1 (avg)
DT	0.7127	0.6018	0.7670	0.5581	0.7975	0.5791	0.7127	0.7101
SVM	0.765	0.7037	0.7986	0.6043	0.8604	0.6502	0.8284	**0.7393**
MultinomialNB	0.7477	0.6335	0.818	0.6821	0.7836	0.6569	0.8005	**0.7287**
LR	0.7357	0.7058	0.7941	0.5911	0.8649	0.7626	0.828	**0.7668**
RF	0.7378	0.791	0.7275	0.3513	0.9497	0.4869	0.8239	0.6554
Vote	**0.7727**	**0.7069**	**0.8018**	**0.6120**	**0.8608**	**0.0.6561**	**0.7727**	**0.7686**

After doing several experiments with each classifier independently, we came up with LR, MultinomialNB and SVM classifiers. In order to improve the performance of each classifier, we choose the best optimization of the parameters in each of them. For the first LR classifier we use the parameter penalty equal to l1 and for the SVM classifier we use the linear kernel.

After seeing the results in Table 2, our last classification model based on *Vote* ensemble classifier combined three individual algorithms: *Logistic Regression (LR)*, *Multinomial Naive Bayes (MultinomialNB)* and *Support Vector Machines (SVMs)*. We have also tested with other models such as *Decision Tree (DT)* and *Random Forest (RF)* 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 training set and we evaluated it with the test set.

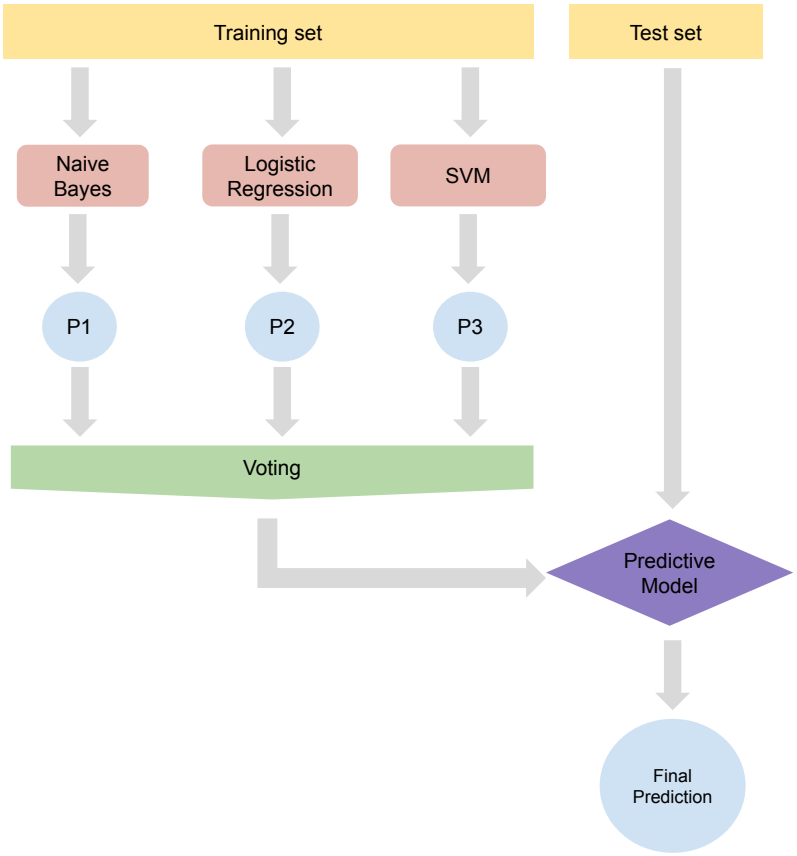

Fig. 1: System architecture.

4 Analysis of results

The system has been evaluated using the official competition metric, the macro-averaged F1-score. It has been computed as follows:

$$\text{Macro-F1} = \frac{2 * \text{Macro-Prec} * \text{Macro-Rec}}{\text{Macro-Prec} + \text{Macro-Rec}} \quad (1)$$

The results of our participation in subtask MEX-A3T of IberLEF Workshop during the evaluation phase can be seen in Table 3.

Table 3: System Results per participating team in subtask MEX-A3T of IberLEF Workshop.

User name (ranking)	F1	F0	Macro-F1
INGEOTEC(1)	0,4796	0,8131	0,6464
Casavantes (2)	0,4790	0,8164	0,6477
Baseline (Trigrams) (9)	0,4300	0,7860	0,6080
Baseline (BoW) (17)	0,3690	0,7830	0,5760
mdmolina (21)	**0,2990**	**0,8232**	**0,5611**
Aspie96 primary (26)	0,2682	0,7939	0,5311

In relation to our results, it should be noted that we achieve better score in case of the class Non AGG (F1: 0.8232). However, our system is not able to classify well the AG class (F1: 0.299).

With respect to other users, we were ranked in the 21th position as can be seen in Table 3.

5 Conclusions and Future Work

In this paper, we describe our participation in IberLEF-2019: subtask MEX-A3T: Authorship and aggressiveness analysis in Twitter: case study in Mexican Spanish [2]. To carry out the task, our classification model is based on Vote ensemble classifier combined three individual algorithms.

For the machine learning approach, we have studied several supervised classifiers: Decision Tree, Support Vector Machine, Multinomial Naive Bayes, Random Forest and Logistic Regression, and the use of n-grams features. It has been observed that when we apply as feature the combination of unigrams and bigrams the Macro F1-score increases in all classifiers. Taking into account the three best classifiers studied, we have combined them via a majority voting ensemble classifier.

In conclusion, we consider that the automatic detection of aggressive language in textual information in general, and in social media in particular, is a very interesting and challenging problem. Besides, we should add the problem of

the different languages and variety of dialects that the Spanish language has, for example, Mexican or Colombian Spanish. Thus, much work needs to be done before an accurate system is finally achieved. Therefore, we will continue studying the problem for different tasks related to hate speech and languages. In particular, since the studies concentrating on Spanish are scarce, we will continue developing systems for detecting hate speech in Spanish and its dialects, as it is one of the most widely spoken languages in the world.

Acknowledgments

This work has been partially supported by Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) and LIVING-LANG project (RTI2018-094653-B-C21)from the Spanish Government.

References

1. Álvarez-Carmona, M.Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H.J., Villasenor-Pineda, L., Reyes-Meza, V., Rico-Sulayes, A.: 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. vol. 6 (2018)
2. Aragón, M.E., Álvarez-Carmona, M.Á., Montes-y Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Moctezuma, D.: Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In: Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain, September (2019)
3. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., Sanguinetti, M.: 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 (2019)
4. Fersini, E., Nozza, D., Rosso, P.: 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 (2018)
5. Fersini, E., Rosso, P., Anzovino, M.: Overview of the task on automatic misogyny identification at ibereval 2018 (2018)
6. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR) **51**(4), 85 (2018)
7. Kumar, R., Ojha, A.K., Zampieri, M., Malmasi, S.: Proceedings of the first workshop on trolling, aggression and cyberbullying (trac-2018). In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018) (2018)
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)

9. Waseem, Z., Chung, W.H.K., Hovy, D., Tetreault, J.: Proceedings of the first workshop on abusive language online. In: Proceedings of the First Workshop on Abusive Language Online (2017)
10. Wiegand, M., Siegel, M., Ruppenhofer, J.: Overview of the germeval 2018 shared task on the identification of offensive language. In: 14th Conference on Natural Language Processing KONVENS 2018 (2018)