Attribute Selection Techniques for Classification of Aggressive Tweets

LyR-UAMC participation at MexA3T 2019 Task

Gabriela Ramírez-de-la-Rosa, Esaú Villatoro-Tello, and
Héctor Jiménez-Salazar

Language and Reasoning Research Group,
Information Technologies Department,
Universidad Autónoma Metropolitana Unidad Cuajimalpa, Mexico City, Mexico.
{gramirez, evillatoro, hjimenez}@correo.cua.uam.mx

Abstract. This paper describes the participation of our Research Group in the shared task of MexA3T 2019. We evaluated the impact of using as principal features, the set of terms identified as discriminant by distinct feature selection strategies. Our main goal was to test if a condensed set of words can be indicative of the aggressiveness of a short text written in a very informal setting (i.e., tweets). Our experiments indicate that different feature selection techniques favor different aspects of the aggressiveness in a short text.

Keywords: Feature selection · Aggressiveness identification · Text classification.

1 Introduction

Hate speech, harassment, and cyberbullying are a few examples of how social media can negatively affect groups of people online. According to [4], aggression in social media is targeted to a particular person or group aiming to damage their identity or lowering their prestige. Previous research proposes that aggression is often expressed in two ways: directly expressed or hidden in the posts [5], resulting in a very challenging task to be performed automatically. Additionally, aggressiveness depends heavily on cultural aspects, as well as the local context, hence, the necessity of building automatic systems for languages different than English, and culturally oriented is becoming more relevant.

Accordingly, in this paper, we describe our proposed system for aggressiveness identification in Mexican-Spanish tweets. Specifically, we describe our participation in the second edition of the MexA3T 2019 challenge [1]. For this particular task, we were given a set of 7700 training tweets, labeled as being aggressive

and non-aggressive. Our approach replicates the traditional pipeline of a non-thematic text classification system. However, contrastingly to previous research using the same dataset, our main goal was to evaluate the impact of distinct feature selection strategies. Thus, we evaluated if a highly condensed set of words are capable of providing to the learning algorithms with valuable and discriminant information solving the posed task. An initial analysis of the obtained features indicates that a distinct aspect of the aggressiveness can be detected by each of the feature selection strategies.

2 System description

Our main goal was to evaluate if a condensed set of terms, obtained through a feature selection strategy, would provide enough information to identify aggressiveness in a tweet. To that purpose, we used the traditional pipeline for text classification. First step is the preprocessing, we lowercased all texts, removed numbers and symbols (except for exclamation points), we used a common word to indicate a hashtag, and all emojis where described[1] in text format. Second, we extracted a set of attributes to generate a representation based on the vector model. In this phase we tested the three strategies described below for selecting a relative small set of attributes. Third, we use a learning algorithm to classify tweets into two classes: aggressive and non-aggressive.

Next, we describe three different strategies for attribute selection. Two of them are widely used in many text classification task, however the third one, to the best of our knowledge, has not been used as feature selection technique before.

2.1 Document frequency

Document frequency is a very simple approach for selecting informative terms, aiming at lowering the dimensionality of vector representations. The idea behind this technique is to ignore terms that are used in very few documents (hapax) or used in almost all documents (known as corpus-specific stop words), assuming its importance is locally to only those texts or they are not useful for distinguishing between classes because those terms appeared in all documents, respectively; hence, we ignore terms not fitting on one of this criteria.

For our experiments we used a grid search to find the best value for these thresholds. We search for both high and low thresholds (high indicate we ignore terms appearing in more than $h\%$ of all documents; and low indicate that we ignore terms appearing in less than $l\%$ of all documents). The best empirical result was obtained when $h = 80\%$ and $l = 0\%$.

2.2 Mutual information

From information theory, mutual information (MI) of two random variables x and y can be describe as the reduction in the uncertainty of variable x due to

[1] We used an emoji library for python: https://github.com/carpedm20/emoji/

the knowledge of y [3, 6]. In other words, MI measures the relevance of a feature x to predict the class y.

Thus, we computed the MI value for each feature, and those with the greater MI value were used as attributes. For our experiments we selected n top features with greater MI, the best performance was obtained with $n = 1000$.

2.3 Lexical availability

This technique is a linguistically motivated approach aiming to identify those lexical markers that represent the words springing to mind in response to a specific topic. Lexical Availability (LA) measures the ease with which a word is generated in a given communicative situation, and allows to obtain the *mental lexicon* which represents the vocabulary flow usable of a person [2]. The terms with greater LA can be seen as the most important ones for a set of tweets from the same class. Thus, we computed the mental lexicon for each class, and then, we used the resulting set of combining both lexicons as features.

In our experiments, for combining the mental lexicon for each class we used the union, intersection and symmetric difference of both lists. However, the best performance was obtained when we used the union of both lexicons to generate the final vocabulary. We also search for the n top attributes with the higher value of lexical availability before combining the lists. The best performance was obtained with $n = 1000$ and $n = 3000$.

3 Results

During the development phase of the MexA3T challenge, a training set of tweets was released. The total number of instances for the aggressive class was 2727, and 4973 for no-aggressive tweets, for a total of 7700 tweets in the training set. A better description of the data can be found in [1].

For validating our experiments in this stage we performed a 10 fold cross validation technique. We used the F-score to measure the best performances; although we use the macro average score we also reported the F-score for the positive class (aggressive) and for the negative class (non-aggressive).

Furthermore, we used several classifiers along the development phase but we found consistent performances with Naïve Bayes classifier. Thus, for all of our reported results we used the *scikit-learn*[2] implementation of this classifier with the default parameters.

Table 1, under the *validation phase* column, shows the best performance for all the configurations tested, using the three different feature selection techniques with some variations as described in Section 2. We also included the bag-of-word (BOW) with the top 1000 most frequent features. As we can see from the table, all feature selection strategies showed similar performances. A slighter better F-macro was obtained with *document frequency* (DF) and *lexical availability* (LA

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

with n=3000). However, note that while DF performed similar to LA (n=3000) the size of the vocabulary for DF was approximately 15000 for each fold; while the vocabulary size for LA with n=3000 was approximately 4500 for each fold.

	Validation phase			***Test phase***			
	F+	F-	Fm	Run-ID	F+	F-	Fm
BOW (n=1000)	0.63	0.75	0.69	LyR_run1	0.37	0.75	0.56
DF (h=80%)	0.63	0.78	**0.71**	LyR_run3	0.42	0.78	**0.60**
MI (n=1000)	0.62	0.76	0.69	LyR_run2	0.38	0.76	0.54
LA (n=1000)	0.63	0.76	0.70	LyR_run4	0.28	0.81	0.57
LA (n=3000)	0.63	0.78	**0.71**	LyR_run5	0.38	0.78	**0.59**
Ensamble	-	-	-	LyR_run6	0.42	0.76	0.58
BOW (baseline-given by track organizers)					0.36	0.78	0.57
Average (all submissions)					0.38	0.76	0.57
Best system (in the task [1])					0.47	0.81	0.64

Table 1. Results in validation and test phases reported in F-score for aggressive (F+), non-aggressive (F-) and a macro average of F-score (Fm). The first column shows the feature selection method used. The parameter used in each row is show in parenthesis (n correspond to the n-top most ranked terms).

For the evaluation stage of the MexA3T shared task a set of 3156 tweets were given. The final results are shown in Table 1 under the *test phase* column. For this stage, we also included an ensemble configuration, which assigns the class corresponding to the majority vote of the five submissions. As we can see, similar as in the validation stage, we obtained the best f-macro with *document frequency* (DF) and *lexical availability* (LA). However, the performance for the positive class (aggressive) decreased, for all submissions, in about 30% in comparison to the validation data.

To have a better idea of the amount of unique information given by the three feature selection strategies, Table 2 indicates the size of unique sets of attributes when comparing pairs of strategies. To make a fair judgment, we examined sets with the same sizes (i.e. 1000). The last row in this table shows, on the one hand, that *lexical availability* (LA) has 191 unique attributes compared with the set obtained by the *bag-of-words* (BOW) vocabulary. On the other hand, only 109 attributes are unique in the set of BOW compared against the LA method. Similarly, when LA is compared against the MI technique, the total number of unique attributes is 212 for LA but only 131 for MI. This initial analysis illustrates that the LA method is capable of finding more diverse attributes in general.

(-)	BOW	MI	LA
BOW	0	178	109
MI	179	0	131
LA	**191**	**212**	0

Table 2. Number of unique attributes when comparing two feature selection techniques. For obtaining these values, we computed the sets difference between the features obtained using the method in the row against the method in the column.

4 Conclusion and future work

This paper describes our participation in the shared task of MexA3T 2019 for identify aggressive tweets written in Mexican-Spanish. We used the general pipeline for text classification varying the strategy used for feature selection. Our goal was to test if a condensed set of words could be indicative of the aggressiveness of a short text. Our experiments indicated that different feature selection techniques favor different aspects of the aggressiveness in a short text.

As feature work we plan to performed a deeper analysis of the set of attributes selected by the tested strategies. Our initial analysis indicates that lexical availability can extract set of attributes with a different linguistic meaning as opposed to bag-of-words or mutual information.

Acknowledgements.
We thank UAM Cuajimalpa and CONACyT (project grant CB-2015-01-258588) for their support. The first author thanks INAOE for the facilities given in her research visit.

References

1. Mario Ezra Aragón, Miguel Á Álvarez-Carmona, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Daniela Moctezuma. 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.
2. Rosa María Jiménez Catalán. *Lexical availability in English and Spanish as a second language*, volume 17. Springer, 2013.
3. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 1991.
4. Jonathan Culpeper. *Impoliteness: Using language to cause offence*, volume 28. Cambridge University Press, 2011.
5. Sreekanth Madisetty and Maunendra Sankar Desarkar. Aggression detection in social media using deep neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 120–127, 2018.
6. Brian C Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014.