Aggressiveness Identification in Twitter at IberLEF2019: Frequency Analysis Interpolation for Aggressiveness Identification

Òscar Garibo i Orts[1][0000−0001−8089−1904]

Universitat Politècnica de València / 46025 València Spain
osgaor@alumni.upv.es

Abstract. This document describes a text change of representation approach to the task of Aggressiveness Identification in Twitter , as part of IberLEF2019. The task consists in classifying tweets as being aggressive or not aggressive. Tweets have been written by Mexican authors who come from a wide variety of backgrounds. Our approach consists of a change of the space of representation of text into statistical descriptors which characterize the text. In addition, dimensional reduction is performed to 6 characteristics per class in order to make the method suitable for a Big Data environment. Frequency Analysis Interpolation (FAI) is the approach we use to achieve rank 12th among 24 submissions.

Keywords: Agresiveness detection · FAI · Author Profiling

1 Introduction

Social media has become a new standard of communications in the last years. Every year more and more people actively participate in the content creation, sometimes under the shield of anonymity. Social media has become a complex communication channel in which usually offensive contents are written. Supervising the content and banning offensive messages currently is a subject of high interest for social media administrators. In this task we address the problem of detecting aggressive comments in tweets from Mexican users. Spanish is a language plenty with nuances, a characteristic which excels in Mexican Spanish and their usage of "albur", where nothing means what it seems. This problem will be considered as an Author Profiling task, since the main goal is building a system which would ideally detect author whose content is offensive to women and/or immigrant.

Author Profiling is widely studied and some new ideas arise from time to time [1]. We have developed a new representation method for text that reduces the

Training	Test
7,700	3,156

Table 1. Number of tweets per dataset.

Class	Training	Test
Non aggressive	4,973	2,372
Aggressive	2,727	784

Table 2. Mexican aggressiveness corpus: distribution of the classes.

dimensionality of the information for each author to 6 characteristics per class. This representation, Frequency Analysis Interpolation, is used to codify the texts for each user and this codified information is used as input data to support vector machines with linear kernel. In a Big Data environment, reducing the number of characteristics from thousands to 6 per class allows an efficient way to deal with high volumes at high speed. With this will in mind a previous method was tested which can be checked at [2] and [3].

2 Corpus

Whereas a complete description of the corpus used in this task can be found at [4], we will have a glimpse and introductory description of basic information in regards of it. The data set for this track consists of tweets that were collected based on their content. Aggressive "Mexicanism" words were explicitly looked for and manually labelled by two people as aggressive or non-aggressive. A tweets was considered aggressive if it contained at least one of the referred words and had the intention to disparage or humiliate a person or a group of people, either by using nicknames, jokes, derogatory adjectives or profanities. In Table 1 we show the number of tweets per dataset, and in Table 2 we show the classes distribution for both datasets.
For this task, the final score corresponds to the F 1 -measure for the aggressive class.

3 Methodology

Our goal was to develop a method that was language independent and that required no prior knowledge of the language used by the authors. We started implementing Term Frequency (TF) representation for each tweet in the corpus, counting how many times each word appears in each author, each tweet in this case, and globally for all tweets. We denote $\mathrm{TF_a}$ as the term frequency vector for author a.

$$TF_{\mathrm{a}} = \left[TF_{(w_1,a)}, TF_{(w_2,a)}, \ldots, TF_{(w_m,a)}\right] \quad (1)$$

TF is used since this way we could represent a priori class dependent probability for each term for each class simply by counting the number of times a term occurs for each class, and dividing this amount by the number of times this term shows for all classes. Let F be the frequency term vector for all classes.

$$F = \sum_{a \in A} TF_{\mathrm{a}} \tag{2}$$

In order to achieve that, one vector per class is generated. The vector length is the number of words in the vocabulary. For each word, we divide the number of times this word shows for this class, and divide it by the number of times the word shows in all classes. We denote $\mathrm{C_k}$ as the term frequency vector for class k that belong to the set of all classes K.

$$C_{\mathrm{k}} = \sum_{a \in A_{\mathrm{k}}} TF_{\mathrm{a}} \forall k \in K \tag{3}$$

These vectors are then used to codify the texts. Each word in the text is substituted by the a priori probability for each class in as many arrays as classes.
Once we have codified the text, six statistic values are calculated for each of the classes:

1. Mean.
2. Standard Deviation.
3. Skewness.
4. First Tertile's length.
5. Second Tertile's length.
6. Third Tertile's length.

At this point, for every author, 6 characteristics per class are calculated and concatenated in a single vector. This vector is used to feed the Support Vector Machines with Linear kernel. LinearSVC support vector machine from Pythons Sklearn library is used to train the model and, of course, to predict the results. One vector is created for each author. This vector contains the six characteristic mentioned above for every class, concatenated.
Although the FAI representation was developed and mainly tested for Author Profiling tasks, it has previously been used for agressiveness detection at HatEval in SemeEval 2019 with good results for multi-class classification [3].

4 Evaluation results

This task is evaluated and ranked using F1-score for the aggressive class.

$$Acc = \frac{\text{number of correctly predicted instances}}{\text{total number of instances}} \tag{4}$$

$$P = \frac{\text{number of correctly predicted instances}}{\text{number of predicted labels}} \tag{5}$$

Team	F-1 Agg. Class
INGEOTEC_task_aggressiveness_run_1	0.4796
Casavantes_Aggressiveness_Text	0.4790
GLP-run2_Aggressiveness_Text	0.4749
GLP-run4_Aggressiveness_Text	0.4635
mineriaUNAM_aggressiveness_secondaryRun	0.4549
mineriaUNAM_aggressiveness_primaryRun	0.4516
GLP-run3_Aggressiveness_Text	0.4405
GLP-run1_Aggressiveness_Text	0.4405
Baseline (Trigrams)	0.4300
LyR$_{A}ggressiveness_Text_Run3$	0.4288
LyR$_{A}ggressiveness_Text_Run6$	0.4212
Victor_run1	0.4081
OscarGaribo_run1	0.3956
LyR_Aggressiveness_Text_Run5	0,3819
LyR_Aggressiveness_Text_Run2	0,3807
LyR_Aggressiveness_Text_Run1	0,3761
Baseline (BoW)	0,3690
OscarGaribo_run2	0,3685
LASTUS-UPF_run2	0,3229
LASTUS-UPF_run1	0,2994
mdmolina_agressive_detection	0,2990
Victor_run2	0,2921
Aspie96_secondary	0,2906
LyR_Aggressiveness_Text_Run4	0,2835
hzegheru_Aggressiveness_Text	0,2786
Aspie96_primary	0,2682

Table 3. Aggressiveness detection task classification.

$$R = \frac{\text{number of correctly predicted instances}}{\text{number of labels in the gold standard}} \tag{6}$$

$$F1 = \frac{2 * \text{P} * \text{R}}{\text{P} + \text{R}} \tag{7}$$

FAI is usually penalized by the fact of having two classes. As we could see at [3] it performs better with a multi-class problem. Nevertheless, as we can see at Table 3, our method has ranked in the middle of the rank table and has overcome BoW Baseline. Since the change of representation depends on the vocabulary that is used, subtle sentences which can denote hate in the speech but which are not using explicit offensive vocabulary might have been mislabeled. For example, polysemic words can be causing mislabelling, since FAI only considers the per class term frequency, but no context is taken into account.

5 Conclusions and future work

We have used FAI, a method developed under the scope of Author Profiling tasks to approach HatEval Task. Prior testing performed with our method has been done under different conditions, since there were always more tweets (minimum 100) per author. Thus, there was much more vocabulary to learn from, and more vocabulary per author. We have to point that our method can easily be updated with new data, since the only required task to be done is recomputing the a priori probability vectors once the new labeled data is available, and train the machine learning algorithm, support vector machines in this specific case. As future work we think of exploring new configurations of our method. One of the immediate ones is to remove some of the vocabulary from the vocabulary we use to codify the tweets. We have seen in our in house testing that some problems require the more the better vocabulary, for example age identification, whereas some others work better if low used words are removed from the vocabulary, for example removing words used by less than 1% of the authors.

References

1. Francisco Rangel, Marc Franco-Salvador, and Paolo Rosso. 2016. A low dimensionality representation for language variety representation. In: Linguistics and Intelligent Text Processing, CICLing-2016, Springer-Verlag, Revised Selected Papers, Part II, LNCS(9624) , pages 156169.
2. Oscar Garibo. 2018. A big data approach to gender classification in twitter. In CLEF 2018 Labs and Workshops. Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org/Vol-2125/paper 204.pdf.
3. 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.
4. Mario Ezra Aragón, Miguel Á Álvarez-Carmona, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda and Daniela Moctezuma. 2019. Overview of MEX-3AT 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.