Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter at SemEval-2019 Task 5: Frequency Analysis Interpolation for Hate in Speech Detection

Òscar Garibo i Orts[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 Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter, as part of SemEval-2019[1]. 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.

1 Introduction.

Author Profiling task is widely studied and some new ideas arise from time to time (Rangel, 2016). We have developed a new representation method for text that reduces the 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 (Garibo, 2018)

2 Corpus

Two corpus have been provided for this task, one in each of the 2 different languages which are subject of study (i.e. English and Spanish). For each language, a training and an evaluation dataset have been provided. The contents of both datasets are individual tweets, that have been collected and manually annotated.

[1] alt.qcri.org/semeval2019/

	Language	Training	Evaluation
Number of tweets	English	9,000	1,000
	Spanish	4,500	500

Table 1: Number of tweets per dataset.

	Language	Training	Evaluation
HS	English	3,783	113
	Spanish	1,857	222

Table 2: Number of Hate tweets per language for Task A.

The goal of this task is to identify tweets which contain hate against women and immigrants. The task has two related subtasks:

1. TASK A. Hate Speech Detection against Immigrants and Women: a two class classification where systems have to predict whether a tweet with a given target (women or immigrants) is hateful or not hateful. This is labeled as a 1 in HS column.

2. TASK B: Agressive Behaviour and Target Classification: where systems are asked first to classify hateful tweets (e.g., tweets where Hate Speech against women or immigrants has been identified) as agressive or not agressive, labeled as AG column in the datasets, and second to identify the target harassed as individual or generic (i.e. single human or group), labeled as TR column in the datasets.

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 TF representation for each tweet in the corpus, counting how many

Language	Class	Training	Evaluation
English	Only HS	1,350	113
	HS and AG	1,092	95
	HS and TR	874	110
	HS, AG and TR	467	109
Spanish	Only HS	279	36
	HS and AG	449	49
	HS and TR	76	10
	HS, AG and TR	1,053	127

Table 3: Number of tweets per case for Task B.

HS	AG	TR	Label
0	0	0	000
1	0	0	100
1	0	1	101
1	1	0	110
1	1	1	111

Table 4: Labels for the SVM.

TASK	English	Spanish
A	63 out of 70	29 out of 39
B	9 out of 42	5 out of 24

Table 5: Our results in Tasks A and B.

times each word appears in each author, each tweet in this case, and globally for all tweets. 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. 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. 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, some statistic values are calculated for each of the classes:

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

These are the six characteristics which are used to feed the Support Vector Machines with Linear kernel. LinearSVC support vector machine from Python's Sklearn library is used to train the model and, of course, to predict the results. In order to provide with the labels for the support vector machines to learn the different labels were concatenated to build a 5 class classifier. In Table 4 the 5 classes which were provided to the support vector machine are shown.

4 Evaluation results.

Our method has performed better in TASK B than Task A. 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. On the other hand, once we provide with more refined labeling, the method tends to catch better the use of aggressive language. This can be seen in the results provided by the task organization. In both languages, English and Spanish our method has obtained better recall values in Task B than A.

5 Conclusions and future work.

Our method has shown to get better results for multi-class classification in the context of this task. Prior testing performed with our method has been done under different environment, since there were always lots of 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 probabilities 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. Since only the last submission was evaluated we still do not know if we can go any further and do better with simple adjustments. 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

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*$_2$04.*pdf*.

Franco-Salvador Rangel, Rosso. 2016. A low dimensionality representation for language variety representation. In *Postproc. 17th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2016, Springer-Verlag, Revised Selected Papers, Part II, LNCS(9624)*, pages 156–169.