OscarGaribo at SemEval-2019 Task 5: Frequency Analysis Interpolation for Hate in Speech Detection

Òscar Garibo i Orts
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]. The task is divided in two sub-tasks. Sub-task A consists in classifying tweets as being hateful or not hateful, whereas sub-task B requires fine tuning the classification by classifying the hateful tweets as being directed to single individuals or generic, if the tweet is aggressive or not. 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 5th in Spanish language and 9th in English language in sub-task B in both cases.

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. Offensive speech can be addressed to individuals or groups due to the race, sexuality, religion and some other characteristics. In this task two of these characteristics will be used as target for offensive speech, women and immigrants. 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 (Rangel et al., 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 corpora have been created to be used in SemEval Task5, HatEval (Basile et al., 2019). 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 datasets have been provided. The contents of both datasets are individual tweets, that have been collected and manually annotated.

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

[1] alt.qcri.org/semeval2019/

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 460–463
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

Language	Training	Evaluation
English	10,000	3,000
Spanish	5,000	1,600

Table 1: Number of tweets per dataset.

Language	Training	Evaluation
English	4,210	1,260
Spanish	2,790	660

Table 2: Number of Hate tweets per language.

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 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 TF_{a} as the term frequency vector for author a.

$$TF_{\text{a}} = TF_{(w_1,a)}, TF_{(w_2,a)}, \ldots, TF_{(w_m,a)} \quad (1)$$

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.

$$F = \sum_{a \in A} TF_{\text{a}} \quad (2)$$

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 C_{k} as the term frequency vector for class k that belong to the set of all classes K.

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 3: Labels for the SVM.

$$C_{\text{k}} = \sum_{a \in A_{\text{k}}} TF_{\text{a}} \forall k \in K \quad (3)$$

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. 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. The same encoding procedure has to be performed for the test dataset. One vector is created for each author. This vector contains the six characteristic mentioned above for every class, concatenated.

4 Evaluation results

Task A is evaluated using standard evaluation metrics, including accuracy (Acc), precision (P), recall (R) and F1-score (F1), while submissions were ranked by F1-score. The metrics were computed as follows:

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

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

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

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

FAI has not achieved great results for 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. For example, polysemic words can be causing mislabelling, since FAI only considers the per class term frequency, but no context is taken into account. Because the method is language independent, the differences of performance between both languages (English and Spanish) depends on the term frequency for each class observed for the train dataset.

Task B is evaluated using the Exact Match measure where all the dimensions to be predicted will be jointly considered computing the Exact Match Ratio . Given the multi-label dataset consisting of n multi-label samples (xi,Yi), where xi denotes the i-th instance and Yi represents the corresponding set of labels to be predicted (HS 0,1, TR 0,1 and AG 0,1), the Exact Match Ratio (EMR) will be computed as follows:

$$EMR = \frac{1}{n}\sum_{i=1}^{n} I(Yi, Zi) \tag{8}$$

Where Zi denotes the set of labels predicted for the i-th instance and I is the indicator function.

Our method has performed better in TASK B than Task A. 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 for both languages for Task B in tables 5 (English) and 6 (Spanish).

As in Task A, the difference of performance of FAI for Spanish and English datasets depends on the term frequency for all classes. Different results have to be expected for different languages.

Ranking	Participant	EM
	MFC Baseline	0.58
1	ninab	0.57
2	iqaameer133	0.568
3	scmhl5	0.483
4	garain	0.482
5	gertner	0.399
6	amontejo	0.384
7	alonzorz	0.382
8	saagie	0.374
9	OscarGaribo	0.373
⋮	⋮	⋮
	SVC Baseline	0.308
⋮	⋮	⋮
42	abaruah	0.159

Table 4: Task B classification for English language.

Ranking	Participant	EM
1	hammad.fahim57	0.705
2	iqaameer133	0.675
3	gertner	0.671
4	francolq2	0.657
5	OscarGaribo	0.6449
6	kwinter	0.638
⋮	⋮	⋮
12	choal	0.616
	SVC Baseline	0.605
⋮	⋮	⋮
16	Taha	0.593
	MFC Baseline	0.588
⋮	⋮	⋮
24	guzimanis	0.428

Table 5: Task B classification for Spanish language.

5 Conclusions and future work

We have used FAI, a method developed under the scope of Author Profiling tasks to approach HatEval Task. FAI 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 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. 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

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.

Òscar 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*.

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 156–169.