HIGHER SCHOOL OF ECONOMICS, MOSCOW

DATA ANALYSIS AND VISUALIZATION

Project Title

Author:

Andrea Grillandi

Professors:

I. Shurov, O. Lyashevskaya

Academic year: 2020/2021

Contents

List of Figures

Chapter 1

Preliminary version of the project

For simplicity I will now keep the first three parts of the project which are required for submission all in the same chapter. Future layout may vary.

1.1 Research objectives and hypothesis

Hate speech is a phenomenon strongly associated to the Internet, and the Internet is indeed one of the main places where *hate speech* is expressed and spread. It can be defined as any sort of expression intended to cause offense or harm basing on sexual orientation, political affiliation, religious beliefs or any other kind of social identifier. This is just a partial definition of the phenomenon, more detailed explanations can be found on Rahman (2020).

This research stems from the ideas present in the just cited paper and, accordingly, from the data related to it. Of course, it also originates from a personal interest in real data and in the possibility of extracting possibly relevant information from those pieces of data. Twitter is the perfect pool from which to fish this kind of data. Thousands of people post on Twitter everyday, expressing their opinion on this or that topic. From these little tweets, we might be able to understand the inclination of person, to sense their sentiment towards a particular topic, or person.

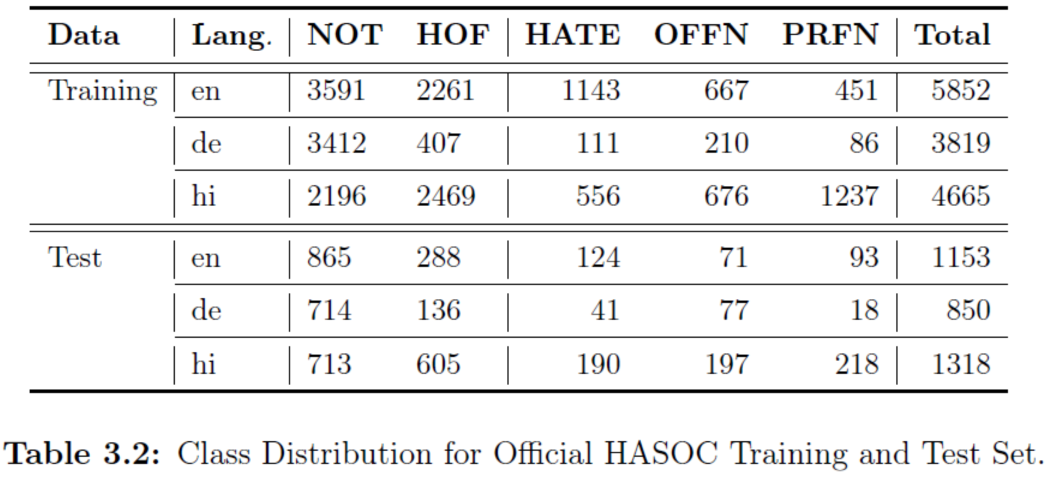
My idea is extracting a bunch (the quantity is yet to be defined) of tweets which contain one or more keywords, using the Twitter API for R. Consequentially, the research would test whether there exists a correlation between the presence of a given keyword and the possible detection of *hate speech* in a text (a tweet, in this specific case). It is thus fundamental to define the nature of these keywords. The idea, so far, is experimenting with names of politicians who are relevant in the American political scene. For example, my keyword list would include names such as Trump, Obama, Clinton, Biden, Sanders, etc...

To conclude, the main aim of this research is testing whether there is or not a correlation between the presence of a given keyword and *hate speech* in a text.

1.2 Description of Input Data

Data will be taken from the following website:
https://hasocfire.github.io/hasoc/2019/dataset.html

According to the already quoted paper, tweets were extracted using a list of keywords or hashtags that included hateful content. They divided their work into two tasks, one is a binary classification task, the other a three-way multi-class classification task, as shown in the table below. In the binary paradigm, data was divided into **Hate and Offensive** (HOF) and **Non Hate-Offensive (NOT)**; the three-way task is meant to classify further the HOF data from the first task into **Hate speech** (HATE), **Offensive** (OFFN), and **Profanity** (PRFN).

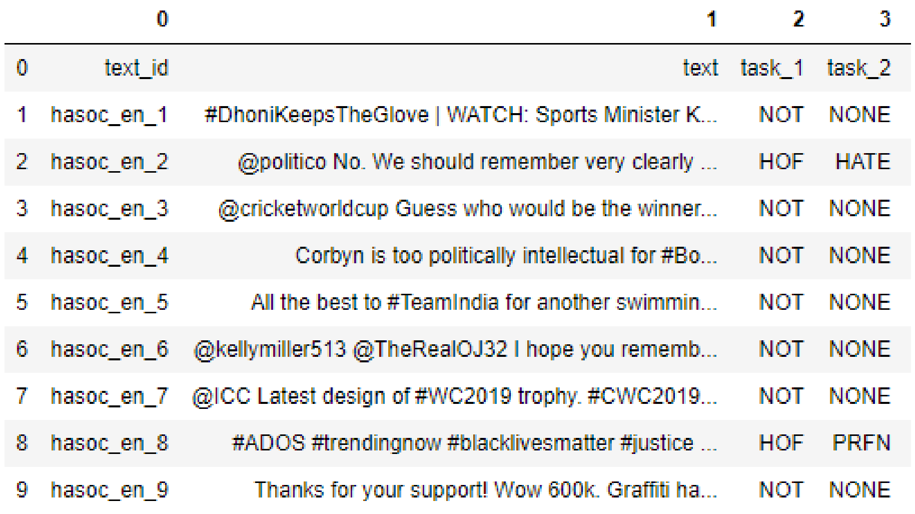


Data	Lang.	NOT	HOF	HATE	OFFN	PRFN	Total
Training	en	3591	2261	1143	667	451	5852
	de	3412	407	111	210	86	3819
	hi	2196	2469	556	676	1237	4665
Test	en	865	288	124	71	93	1153
	de	714	136	41	77	18	850
	hi	713	605	190	197	218	1318

Table 3.2: Class Distribution for Official HASOC Training and Test Set.

Figure 1.1: The table was taken from the reference paper Rahman (2020)

Here's how the tsv table of HASOC dataset looks like.



	0	1	2	3
0	text_id	text	task_1	task_2
1	hasoc_en_1	#DhoniKeepsTheGlove | WATCH: Sports Minister K...	NOT	NONE
2	hasoc_en_2	@politico No. We should remember very clearly ...	HOF	HATE
3	hasoc_en_3	@cricketworldcup Guess who would be the winner...	NOT	NONE
4	hasoc_en_4	Corbyn is too politically intellectual for #Bo...	NOT	NONE
5	hasoc_en_5	All the best to #TeamIndia for another swimmin...	NOT	NONE
6	hasoc_en_6	@kellymiller513 @TheRealOJ32 I hope you rememb...	NOT	NONE
7	hasoc_en_7	@ICC Latest design of #WC2019 trophy. #CWC2019...	NOT	NONE
8	hasoc_en_8	#ADOS #trendingnow #blacklivesmatter #justice ...	HOF	PRFN
9	hasoc_en_9	Thanks for your support! Wow 600k. Graffiti ha...	NOT	NONE

Figure 1.2: Table from python pandas library

For the purpose of this research I will be interested only on the results of binary classification task, on which I will train the SVM classifier. A more detailed description of the classifier will follow in the next section.

1.2.1 Twitter data

Further, I will extract data from Twitter (using the rtweet API (https://rtweet.info/)). Extracting data here means the equivalent of carrying out an advanced search where a keyword is used to select the most relevant tweets in a given language, with the option of selecting a time frame the tweets should be from. As stated above, the keyword list will consist of prominent American politician names, such as Trump, Obama, Biden, and so on.

Each of the extracted text will be classified using the classifier we trained on the HASOC dataset. Then, a new classifier will be trained. The presence of every keyword in the text will constitute a feature in the new classifier that will be trained. This way, most probably using a Random Forest classifier, I will be able to evaluate the importance of each feature, thus trying to answer the following: does the presence of e.g. the word "trump" helps us classify a text as being hateful?

1.3 Methods of Analysis

1.3.1 First classifier - SVM

Support Vector Machines (SVM) are machine learning models which are able to perform regressions both on linear and non-linear data. According to Rahman (2020), SVM models are shown to perform better tasks related to hate speech detection. R provides us with a library from David Meyer called *e1071*, containing the *SVM* function, which can be fed with a dataframe containing the required features.

Here's about the feature. First of all, I will preprocess the text of the HASOC dataset tweets by removing stopwords; secondly, I will create a Tf-Idf vector model of the texts (most probably using python scikit-learn); after which I will perform a Byte-Pair Encoding (BPE) of the texts; lastly, I will compute a word n-gram(1, 3) model of the texts. All these abovementioned characteristics will be used as features to train the SVM classifier on the HASOC dataset.

Finally, I will use the computed classifier to classify all the texts extracted from Twitter, then proceed with second phase of the research.

1.3.2 Second Classifier - Random Forest

Random forest is a learning method for classification of quantitative and categorical data that operates by constructing a multitude of decision trees for different batches of data (boosting methods are used), and by eventually averaging the performance of various trees. In R there are different libraries dedicated to it, I will most probably use *randomForest* by Andy Liaw.

For the second part of the research I will train a Random Forest model on the dataset of tweets extracted according to the list of keywords, the content of which was outlined above in the text. However, this time the model will present a series of added features, namely the presence of a given keyword in the text. So, if the text contains a keyword it will present a 1 in the column corresponding to that specific keyword, and, in the same fashion, it will present a 0 in every column corresponding to a keyword which is not present in that text. This way, I want to test the correlation of encountering that keyword in the text and the presence of *hate speech* in it.

Moreover, Random Forest classifiers are able to performance feature importance tests, allowing for a ranging of the variables in test.

Bibliography

Rahman, Md Ataur (2020). "Exploring Features for Multi-label Hate Speech Detection." In: