A Big Data approach to gender classification in Twitter.

Notebook for PAN at CLEF 2018

Òscar Garibo i Orts

Optical Tech & Support
oscar.garibo@ots-es.com

Abstract. This paper describes a statistical approach to the task of gender classification in tweets, with a Big Data perspective in mind. Our task started developing our own implementation of Low Dimension Representation method, with the idea to add some other statistics which had not been used in the original implementation, such as skewness, kurtosis and central moments. Exploratory analysis of the new characteristics showed the importance of skewness due to the problem presents only 2 classes. Our approach will only use skewness for describing the difference in use of the language between men and women and skewness, as well, will be used to predict gender for the test dataset.

1 Introduction

Author Profiling task is being widely studied and some new ideas rise from time to time. We will be pursuing a model to classify tweets that fits into a Big Data environment. Big Data refers not only to huge amounts of data, but data to be processed near real time or at very high rates. About 6,000 tweets are produced per second worldwide [1].

We could consider models based on Deep Learning, which are time and resources expensive to build and train, and to predict as well. Our goal, in the scenario of PAN Author Profiling [1] task was to implement and study the Low Dimension Representation (LDR) [2] approach, since LDR provides a dimension reduction scenario. The possibility of reducing the number of characteristics from thousands to as few as possible was interesting in the context of a Big Data application. Reducing the time and computing resources was our main fuel. We were so keen on that goal that we decided to work with a machine with one core and 4 Gb of RAM, what had to be enough to perform the task.

[1] Internet live stats. http://www.internetlivestats.com/twitter-statistics/

The difference in the usage of the language between men and women has been agreed, basically based upon differences in style, usage or linguistic resources, etc. Tweets present a challenge since so few words are used in each tweet. We tried to find out if despite of the limited amount of words used, we could also establish that men and women use the language in different ways. We also wanted to check if a simple method was good enough at detecting such differences, and thus classifying user's gender.

2 Dataset

The PAN Author Profiling Train corpora is composed by 3 different Tweets collections in different languages (English, Spanish and Arabic). For English and Spanish languages the Twitter Classification Training Dataset contains 3,000 users and for Arabic language it contains 1,500 users. In all three Train Datasets we were provided with 100 tweets per user. Half the users per language are of one sex, thus the other half of the complementary sex.

	English	Spanish	Arabic
Male	1,500	1,500	750
Female	1,500	1,500	750

Table 1

3 Methodology

Our goal was to develop a method that was not language dependent, and that required no prior knowledge of the language.

We started implementing a modified LDR by using Tf representation instead of Tf-Idf and adding new statistical features, such as skewness, kurtosis and central moments.

We decided to try Tf because this way we could represent class dependent a priori probabilities of each term for each class simply by counting the number a term occurs for each class, thus dividing this amount by the total number of times this term shows for all classes. For the sake of a clearer understanding we wanted to be sure we were dealing with probabilities. In addition, calculating Tf is less time and resources consuming than calculating Tf-Idf.

The resulting class dependent arrays are complementary for each term, since we are dealing with just 2 classes.

We then build a vocabulary set including each word we have seen in the train corpus. We decided to discard the words that were appearing less than 5 times in the whole corpus, instead of 5 users as it was originally done in the LDR implementation.

We then went over the train corpus, one user each time, checking the words for this user and writing down in an array the related a priori probabilities. We finally get one vector per user, with the a priori probabilities of each word to pertain to one of the classes.

We then can calculate the different statistics from these a priori probabilities arrays that represent the text used by each user.

Since both arrays were complementary we decided to build the array just for one of the classes, saving time and resources.

It is important to notice that when we want the model to include more tweets to build the a priori probabilities vectors, we only have to run the procedure with new labeled tweets. This new vector is what we will use to predict new incoming tweets. The whole procedure is simple and fast, and can be done in parallel to predicting task. Once the new vector is ready we only need to point the predicting algorithm to point to this new vector.

This is an easy way of keeping the vectors up to date, and more important, this is a clean, fast and reliable updating procedure.

4 Exploratory Analysis

Since we are working with a 2 classes problem, and due to the a priori probabilities we use to encode and represent the text are complementary, the skewness was also complementary for each of the classes. We found out that skewness was certainly helpful. We run the method for all 3 languages for the train corpus. Once we get the skewness in each case we can see that skewness is mostly > 0 for male labeled users, and < 0 for female labeled users.

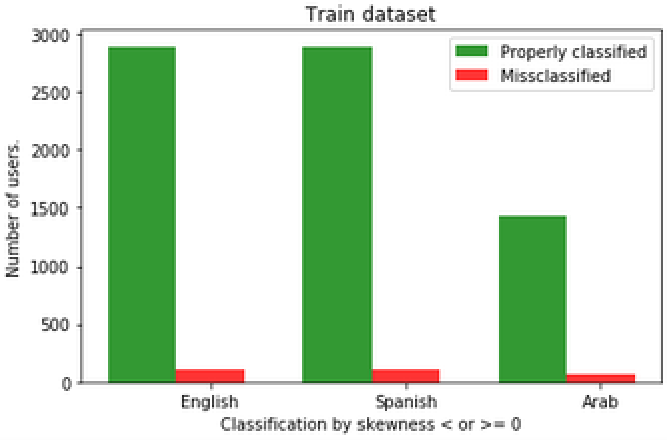

Figure 1.

As it can see in Table 2, more than 95% of the users showed a positive skewness if they were men and negative if they were women. We can think of using this datum to certify men and women use language differently, no matter the language.

Language	Accuracy
Arabic	0.9593
Spanish	0.9647
English	0.9647

Table 2

5 Evaluation Results

We decided to use a very simple method. We are loading the a priori probabilities array for male class (female could also have been used, we just picked one since they are complementary) and use it code the test dataset. We read each user's tweets, go over them codifying the a priori probabilities and then calculate the skewness for the resulting array. If the skewness was positive we then labeled the user as male, and female if it was negative.

Our classifying method ends up with a simple IF. No machine learning was used. We would simply rely on the knowledge we got from the training corpus.

The results are shown in Table 3.

Language	Accuracy
Arabic	0.6750
Spanish	0.7164
English	0.7363

Table 3

Besides accuracy there is another factor we want to focus on, and it is the time consumed in the prediction. We always need to keep in mind we wanted to develop a method capable of working in a near real time approach.

Table 4 shows the time consumed to predict the test datasets, the number of documents in each test dataset (we consider a document all tweets related to one single user), documents processed per second and milliseconds to process each document:

Language	Time to predict	Number of docs	Docs per second	Millisecs per doc
Arabic	2 seconds	1,000	500	2.00
Spanish	3 seconds	2,200	733	1.36
English	3 seconds	1,900	633	1.58

Table 4

We have built a solution that matches the near real time approach in a Big Data environment with thousands of tweets per second arriving to our SW. Whereas the accuracy is not top standing, the processing time fits the tweets volume we could have to handle.

6 Conclusions and Future Work.

We have established that when codifying the text by the words' term frequency using a priori class dependent arrays skewness is highly discriminative for gender classification purposes. This characteristic could be added to any models that could be used in such a task.

We showed that men and women use language in different ways, good enough to reach up to 73% accuracy with a very simplistic approximation.

We also showed that such accuracy can be obtained with very low resources and time consuming approach. Our method could be used in a near real time environment and be used to establish a prior classification. That is, we could certainly establish that a tweet has a 0.75 probability to belong to one of the classes. From that point, we could reduce granularity if finer classification is required.

We might encounter problems in which no 100% classification is required but an approximation on the number of men or women in an environment, for example, what is the % of men in a chat? We could provide with an approximate number by assigning probabilities to each user.

We want to test the skewness approach as a descriptive approximation with as many corpora as possible. This method could not only contribute to improve accuracy as a complement to other approaches, but also as a descriptive method to clarify how men and women differently use language.

Building a priori probability arrays with bigger corpus could help to clarify the foundations of this method.

7 References

[1] Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)

[2] Rangel F., Franco-Salvador M., Rosso P. A Low Dimensionality Representation for Language Variety Identification. In: Postproc. 17th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2016, Springer-Verlag, LNCS(), pp. (arXiv:1705.10754)