Author Profiling for Gender and Language Variety classification in Twitter using Spanish language.

Òscar Garibo i Orts
osgaor@alumni.upv.es

Abstract

In this paper it will be shown that Author Profiling is a mixture of methodology and art. In this work, the provided Baseline is beaten by a minimum of 16,5% using limited resources, time and field specific knowledge.

An average laptop, with an i3 processor and 16Gb RAM together with an average SATA HD, will be used for developing the proposed solution. The equipment will limit the tests that will be done due to lack of memory. Nevertheless, it will still be able to mix statistical methods such as Tf-Idf and Twitter environment features (such as hastags, retweets, etc) counting, and linguistic methods like Part Of Speech tagging.

Three different machine learning methods will be used, Support Vector Machines, Random Forest Classifier and Multi-Layer Perceptron. The latest will offer the best accuracy results.

And all of these has been done in less than three weeks as part of the evaluation of the subject "Text Mining in Social Media" under the scope of the master course Big Data Analytics at Universitat Politècnica de València.

1 Introduction

Author profiling consists of identifying some text's authors based on characteristics found in the text. At a first glance we can think of many practical uses for author profiling such as:

- Marketing. Classifying user by the way they write can be very useful for on-line marketing.
- Security. Author profiling is already being used to detect threats in social media. From terrorist detection to paedophiles detection in social media.

In this task Tweets will be used as the object of study. Twitter can be thought of as a micro-blogging tool. One of its main characteristics is that the number of characters is limited to 140. Twitter has also its own specific vocabulary. Twitter users can create their own contents or can Retweet other's contents. A Twitter user can talk about an specific subject by subscribing a Hashtag. She can also Mention some other Twitter user, and can refer to a URL content. A Twitter user can also use Emojis instead of words. All these, some Twitter specific, other more general, features will be used to improve the predictive performance of our models.

It will be studied if some or any of these characteristics can be used for classifying Tweets, together with language specific features.

As language specific features Tf-Idf, which determines the frequency of each word in relation to the entire corpus, will be used together with a Part Of Speech tagging analysis.

2 Dataset

PAN is a series of scientific events and shared tasks on digital text forensics [1]. Under the development of PAN CLEF 2017 [2] an specific task of author profiling is developed. For such a task, a dataset is provided for participants to build their models to classify Tweets in Variety and Gender. Variety meaning different variations of the same language depending of the country it is spoken, and Gender discerning either the Tweet's author is male or female.

[1] http://pan.webis.de/index.html
[2] http://pan.webis.de/clef17/pan17-web/index.html

The PAN CLEF 2017 Twitter Classification Dataset contains 500 authors per gender and language variety, with 100 classified tweets per author.

For the task under our scope a subset of the original corpus is given. The subset consists of 300 authors per class (gender and language variety), 200 will be used for training and 100 for test, with 100 tweets per author. Variety is composed of 7 different classes, one for a different country where Spanish language is spoken (Spain, M'exico, Colombia, Argentina, Chile, Per'u and Venezuela) while Gender is composed by two classes, classifying either the Tweet's author is male or female. That makes a global number of 14 classes (7 varieties * 2 genders), so 2800 authors with 100 tweets per user make the train dataset, and 1400 authors, again with 100 tweets per author, make the test dataset.

The dataset has been labeled by the PAN organization and all classes are balanced.

The amount of data to be dealt with will not be suggesting to face the task as a Big Data challenge, but when thinking of a classification system, classifying thousands of tweets per minute, then the task definitely lands on the Big Data scene.

3 Student's approach

At class, a Baseline based on a Bag of Words approach was built. Which provided with the following Baseline accuracies:

	Accuracy	
	Variety	Gender
Baseline	77.21	66.43

Table 1: Baseline Accuracy.

Multiple strategies will be considered in order to improve the given Baseline, each of them will be explained:

- Tf-Idf. Tf-Idf sparse matrix will be built for the corpus, considering unigrams, bigrams and trigrams. Corpus specific stop words will be removed by setting a max frequency of 0.90 , which means the 10% most used words will be discarded. The effect of removing unique words by discarding the 5% least used words will as well be explored.

- Several Twitter specific features will be counted:

- Word Count, number of words in tweets per user.
- Emoji Count, number of Emojis per user.
- Hashtag Count, number of Hashtags per user.
- Mentions Count, number of Mentions per user.
- URL Count, number of URLs referred per user.
- Retweet Count, number of Retweets per user.

4 Experimental results

In this task, accuracy will be used as performance measurement to compare the throughput of the solution.

A common step to all the approaches is that the Spanish stop-words have been removed from the corpus prior to any further actions.

In the first approach punctuation marks have removed from the corpus and Support Vector Machines (SVM) have been used to build the classifier. Tf-Idf has been used to measure the relative frequency for each word in each tweets in regards of the corpus. Several configurations have been run to test the model's accuracy.

Tf-Idf has been used with unigrams, bigrams and trigrams. Several max df values have been tested to remove corpus specific stop-words, as well as min df values to remove unique words in the corpus.

As shown in Table 4 the best performance either for Variety and Gender has been achieved with Tf-Idf for unigrams and a max df of 0.90, that is, removing the top 10% most frequent words.

	Accuracy	
SVM no punctuation	Variety	Gender
TfIdf Unigrams	91.21	74.86
TfIdf Bigrams	89.64	73.14
TfIdf Trigrams	89.07	71.86
TfIdf 0.90 Unigrams	**91.29**	**75.07**
TfIdf 0.80 Unigrams	91.14	74.79
TfIdf 0.95 Unigrams	91.21	74.86
TfIdf 0.90 0.05 Unigrams	85.43	74.36

Table 2: SVM + punctuation removal + different max df and min df.

As a second approach, performance performance has been tested while keeping punctuation marks in the corpus. As shown in Table 3 accuracy for Variety remains the same, but some improvement can be seen in Gender classification accuracy.

SVM with punctuation	Accuracy	
	Variety	Gender
TfIdf Unigrams	90.71	**75.71**
TfIdf Bigrams	88.79	73.36
TfIdf 0.90 Unigrams	**91.29**	75.21

Table 3: SVM + keep punctuation + different max df.

As a third approach a Random Forest Classifier has been tested, since during the tests performed in class this meant a huge improvement in Variety accuracy. As shown in Table 4, RF worsens the accuracy both for Variety and Gender. So far, RF will be discarded as a classifier for this task.

RF no puntuation	Accuracy	
	Variety	Gender
TfIdf 0.90 Unigrams	**90.43**	69.14
TfIdf 0.90 Bigrams	87.36	**69.93**
TfIdf 0.90 Trigrams	84.36	68.14

Table 4: RF + punctuation removal + max df 0.90.

As a fourth approach using Multi-Layer Perceptron (MLP) has been introduced to see if performance can be improved using a different machine learning method. MLP has been used with the configuration that offered the best results with SVM. The same configuration that was used for SVM has been used to test MLP, keeping accents and removing them form the corpus. As shown in Table 5 the best results for Variety classification were achieved when using MLP and removing accents. Accuracy for Gender classification was also improved and this configuration will be used as an starting point for further testing.

MLP	Accuracy	
	Variety	Gender
TfIdf 0.90 Unig no accents	**91.79**	76.57
TfIdf 0.90 Unig accents	88.71	**76.64**

Table 5: MLP with and without accents.

How stemming and using Twitter specific features do change the accuracy for Variety classification will be checked. Therefore, some testing were run with the results shown in Table 6.

MLP	Accuracy
TfIdf 0.90 Unigrams no accents stem	91.64
TfIdf 0.90 Unigrams no accents TAGS	91.00

Table 6: MLP + Stem and TAGS.

From the results in Table 8 it can be seen that stemming the words makes no improvement in the accuracy for classifying the Tweets under Variety scope.

At this point Tf-Idf with max df 0.90 and accents being removed from corpus will be kept as the best option for Variety classification.

Further digging on Twitter specific features will be done in order to try to increase the accuracy for Gender classification.

Stanford POS-Tagger was run for both Training and Test datasets. The following features were also built: wordcount, emoji count, hashtag count, retweet count, mention count and URL count, all of them counts per tweet's author, standarized so the counts range between 0 and 1.

Intensive testing results are shown in Table 7, which follows:

- WC = Word Count.
- EC = Emoji Count.
- HC = Hashtag Count.
- MC = Mentions Count.
- UC = URL Count.
- RC = Retweet Count.

MLP	Accuracy
TfIdf 0.90 TAGS + WC	**77.43**
TfIdf 0.90 TAGS + WC + EC	76.07
TfIdf 0.90 Adj + WC + EC	77.36
TfIdf 0.90 TAGS + WC + EC + HC	77.14
TfIdf 0.90 TAGS + EC	76.71
TfIdf 0.90 TAGS + EC + HC	76.93
TfIdf 0.90 TAGS + EC + MC	76.64
TfIdf 0.90 TAGS + EC + UC	76.21
TfIdf 0.90 TAGS + EC + RC	75.79

Table 7: Common:MLP + keep punctuation.

The best results are obtained when performing Tf-Idf with max df of 0.90, together with POS Tags and Word Count. It can be seen that different combinations of Twitter specific features offer a narrow range of accuracies, being selected the ne with higher accuracy.

As a result of all the tests done, an interesting improvement has been achieved in comparison to the provided Baseline. These improvements in Variety and Gender classification are shown in Table 8, which follows.

	Accuracy	
	Variety	Gender
Baseline	77.21	66.43
Our best	**91.78**	**77.43**

Table 8: Accuracy improvement.

According to the official results published at 2017 PAN website [3] our models achieved a notable 10th place out of 18 contenders for Variety classification, and a meritorious 8th place out of 18 contenders for Gender classification.

Multi-Layer Perceptron Model has proved to train and predict faster than SVM by 3 to 4 times. Both models consume a similar amount of memory, but while SVM uses only one of the cores, MLP makes a better usage of the cores, with 2 or 3 of them working simultaneously.

The main conclusions that can be extracted from the work done here are: - Keeping punctuation improves Gender prediction, as well as adding POS tags and tweets word count. - Removing corpus specific stop words by setting a max df value of 0.90 while running Tf-Idf improves both Gender and Variety prediction. Which makes sense, since words used by all categories make no difference for classification. - Removing accents from corpus helps to classify by Variety, since all nationalities seem to use accents a similar way. But it helps classifying Gender. We can assume men and women make a different usage of accents.

5 Conclusions and future work

Something to be remarked is that this task has been done as an exercise for Text Mining in Social Media's subject in the development of the master course in Big Data Analytics at Universitat Politècnica de València. Limited time was available to learn about the subject and to develop legacy code and testing. The teaching team provided with a main structure of code written in R language, that is what was used for preliminary modeling at class. We then decided to start from zero and build our own code in Python for our better convenience. A Baseline built on a Bag of Words was also provided. Baseline was beaten by nearly a 18% for Variety and a 16.5% for Gender.

As future work some actions are proposed in order to improve the results:

- URL can be expanded and checked. This could improve accuracy for Variety classification.
- Specific dictionaries could be used to built extra features that could be of help to improve either Gender and Variety accuracy, depending on the dictionaries built.

[3] http://pan.webis.de/clef17/pan17-web/author-profiling.html