Multilingual Author Profiling using Word Embedding Averages and SVMs

Roy Bayot
School of Electrical and
Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: d11668@alunos.uevora.pt

Teresa Gonalves
Twentieth Century Fox
Springfield, USA
Email: tcg@uevora.pt

***Abstract*—The abstract goes here.**

I. Introduction

Author profiling has been of importance in the recent years. From a forensic standpoint for example, it could be used to determine potential suspects by getting linguistic profiles and identifying characteristics. From a business intelligence perspective, companies could target specific people through online advertising. By knowing the profile of the authors, companies would easily find what a specific group of people talk about online and device strategies to advertise to these people. They could also analyze product reviews and know what types of products are liked or disliked by certain people.

The growth of the internet where text is one of the main forms of communication is one of the reasons for a rising interest in author profiling. Through this growth, various corpora could be extracted, curated, assembled from different sources such as blogs, websites, customer reviews, and even twitter posts. Of course, this presents some problems. For example, people from different countries who use the same online platform such as Twitter or Blogger could behave differently in terms of text usage. This presents a difficulty in profiling. This work tries to take this difficulty into account by studying which kind of features are useful for different languages.

The aim of this work is to investigate the effect of syntactic information on author profiling in different languages. For this purpose, we used the dataset from PAN 2015 [**?**] since it has 4 different languages and profiled almost in the same way on age, gender, and 5 personality traits - agreeability, conscientiousness, extrovertedness, openness, and stability. The four languages are English, Dutch, Italian, and Spanish. There are three different sets of features that are investigated in this work. The first set is the bag of words features in the form of term frequency and term frequency inverse document frequency. The second set are word ngrams. And finally, we also study part of speech ngrams to see if information extracted from these features are useful for characterizing twitter users.

II. Related Literature

In previous author profiling research, most of the work is centered on hand crafted features as well as that which are content-based and style-based. For instance, in the work of Argamon et al. in [**?**] where texts were categorized based on gender, age, native language, and personality, different content-based features and style-based features were used. In another example of Schler et al. in [**?**] where writing styles in blogs are related to age and gender. Stylistic and content features were extracted from 71,000 different blogs and a Multi-Class Real Winnow was used to learn the models to classify the blogs. Stylistic features included parts-of-speech tags, function words, hyperlinks, and non-dictionary words. Content features included word unigrams with high information gain.

This can also be seen in the previous PAN editions. In the first edition of PAN [**?**] in 2013, the task was age and gender profiling for English and Spanish blogs. There were a variety of methods used. One set includes content-based features such as bag of words, named entities, dictionary words, slang words, contractions, sentiment words, and emotion words. Another would be stylistic features such as frequencies, punctuations, POS, HTML use, readability measures, and other various statistics. There are also features that are n-grams based, IR-based, and collocations-based. Named entities, sentiment words, emotion words, and slang, contractions and words with character flooding were also considered. The work of Lopez-Monroy in [**?**] was considered the winner for the task although they placed second for both English and Spanish where they used second order representation based on relationships between documents and profiles. The work of Meina et al. [**?**] used collocations and placed first for English while the work of Santosh et al. in [**?**] worked well with Spanish using POS features.

In PAN 2014 [**?**], the task was profiling authors with text from four different sources - social media, twitter, blogs, and hotel reviews. Most of the approaches used in this edition are similar to the previous year. In [**?**], the method used to represent terms in a space of profiles and then represent the documents in the space of profiles and subprofiles were built using expectation maximization clustering. This is the same method as in 2013 in [?]. In [?], n-grams were used with stopwords, punctuations, and emoticons retained, and then idf count was also used before placed into a classifier. Liblinear logistic regression returned with the best result. In [?], different features were used that were related to length (number of characters, words, sentences), information retrieval (cosine similarity, okapi BM25), and readability (Flesch-Kincaid readability, correctness, style). Another approach is to use term vector model representation as in [?]. For the work of Marquardt et al. in [?], they used a combination of content-based features (MRC, LIWC, sentiments) and stylistic features (readability, html tags, spelling and grammatical error, emoticons, total number of posts, number of capitalized letters number of capitalized words). Classifiers also varied for this edition. There was the use of logistic regression, multinomial Nave Bayes, liblinear, random forests, Support Vector Machines, and decision tables. The method of Lopez-Monroy in [?] gave the best result with an average accuracy of 28.95% on all corpus-types and languages.

In PAN 2015 [?], the task was limited to tweets but expanded to different languages with age and gender classification and a personality dimension. The different languages include English, Spanish, Italian, and Dutch. There were 5 different personality dimensions - extroversion, stability, agreeableness, conscientiousness, and openness. And in this edition, the work of Alvarez-Carmona et al. [?] gave the best results on English, Spanish, and Dutch. Their work used second order profiles as in the previous years as well as LSA. On the other hand, the work of Gonzales-Gallardo et al. [?] gave the better result for Italian. This used stylistic features represented by character n-grams and POS n-grams.

Since the current task is to train on one type of corpus and test on another type of corpus, we decided to try an approach that uses word embeddings. We used word2vec in particular as described in [?] [?]. Such embeddings were trained not on the corpus given by PAN but by Wikipedia dumps so there is a possibility that using such embeddings which work on one corpus type could work on another corpus type. Our approach also uses these embeddings in conjunction with Support Vector Machines.

III. Methodology

The methodology is illustrated by the figure 3. It mainly consists of three parts - word embedding creation, training, and evaluation. These will be further discussed in the subsequent subsections.

A. Word Embeddings Creation

To represent words by a vector, word embeddings have to be created. These vectors capture some semantic information between words. One way to do such embeddings are with word2vec as proposed by Mikolov in [?] and [?]. Essentially, words in a dictionary by a given corpus are initially represented with a vector of random numbers. A word's vector representation is learned by predicting it through its adjacent words. The basis for the order of the words is in a large corpus. This is illustrated in figure 1. The implementation can be two different ways - skip grams and continuous bag of words (CBOW). In CBOW, the word vector is predicted given the context of adjacent words. In skip grams, the context words are predicted given a word.

Fig. 1. Diagram for word2vec implementations
[scale=1]sentiment$_0$1$_l$$arge.png$

Fig. 2. Overview of word2vec flow.
[scale=.5]Word2Vec.png

For our problem, we used wikipedia dumps as an input to the word2vec implementation of gensim [?]. The wikipedia dump used for the following experiments were that of 05-02-2016. As for word2vec parameters, no lemmatization was done, the window size used was 5, and the output dimensions used was 100. The default continuous bag of words was also used. For further details, please refer to the tutorial given in [?].

B. Training and Evaluation

After obtaining word2vec representations for each word as illustrated in figure 2, each xml document of one twitter user is converted into word2vec representations. To do this, the texts were first extracted from the file. Then it was converted to lower case. After the conversion, the words are checked against the dictionary of all the words that have word2vec representations. If the words exists in the dictionary, the vector representation is pulled out and accumulated, and later normalized by the number of words that could be found in the dictionary. If the word does not exist in the dictionary, a zero vector is returned.

After representing each twitter user, the vectors are then used as features. Support Vector Machines were then trained using those features. Different kernels and parameters were also checked. This includes polynomial kernel and a radial basis function. For the polynomial kernel, the degrees were restricted to 1, 2, and 3. The C parameter was restricted to 0.01, 1, 100. For the radial basis function, the gammas and C parameters were restricted to 0.01, 1, 100.

The performance of the system was evaluated using the accuracy measure and 10 fold cross validation was used. The parameters that gave the highest accuracies were noted and used in the system deployed in the TIRA server.

IV. Results and Discussion

The tables I- V give the all the results for English, Spanish, and Dutch on age and gender. Looking at table I for age classification in English, the highest accuracy obtained is 44.8%. The SVM parameter that gave the best classification is the one with the radial basis function kernel with C to be 1 and gamma to be 100 although most of the other values are close. In gender classification however, the highest accuracy obtained

Fig. 3. Overview of the system

[scale=.5]System.png

was 68.2% using a polynomial kernel with the degree to be 3 and C to be 100. There is more variety from these results given that the lowest is around 50.0%.

TABLE I
AGE CLASSIFICATION RESULTS FOR ENGLISH

	poly			rbf		
	degree			gamma		
C	1	2	3	0.01	1	100
0.01	0.418	0.416	0.416	0.414	0.414	0.414
1	0.418	0.416	0.416	0.414	0.418	**0.448**
100	0.418	0.423	0.393	0.416	0.409	0.426

The results for Spanish tweets are given below. In table III, the highest accuracy for age classification is 51.3%. This is given by a classifier with a radial basis function kernel with gamma to be 1 and C to be 100. In table IV, the highest accuracy for gender classification is 67.1%. This was given by the classifier that used a radial basis function kernel with gamma to be 1 and C to be 100.

TABLE II
GENDER CLASSIFICATION RESULTS FOR ENGLISH

	poly			rbf		
	degree			gamma		
C	1	2	3	0.01	1	100
0.01	0.534	0.495	0.495	0.498	0.500	0.512
1	0.534	0.561	0.579	0.498	0.563	0.643
100	0.534	0.677	**0.682**	0.548	0.672	0.643

TABLE III
AGE CLASSIFICATION RESULTS FOR SPANISH

	poly			rbf		
	degree			gamma		
C	1	2	3	0.01	1	100
0.01	0.506	0.506	0.506	0.506	0.506	0.506
1	0.506	0.511	0.511	0.506	0.506	0.496
100	0.506	0.513	0.415	0.506	**0.513**	0.422

Dutch gave the highest accuracy of 71.9% using an SVM with a radial basis function with a gamma of 1 and C of 100. This is further illustrated in table V.

TABLE IV
GENDER CLASSIFICATION RESULTS FOR SPANISH

	poly			rbf		
	degree			gamma		
C	1	2	3	0.01	1	100
0.01	0.504	0.504	0.504	0.504	0.557	0.565
1	0.504	0.546	0.577	0.504	0.573	0.638
100	0.504	0.663	0.654	0.568	**0.671**	0.621

Finally, we also add the last table **??** which shows the results given by PAN after using the classifier on a different corpus type. We can see that there is a drop in accuracy between the one tested on tweets and the one on unknown corpus type. For English age classification, we started with 44.8% which dropped to 35.9%. 62.8%

For Spanish age classification, we started with 51.3% which dropped to 48.2%, which doesnt seem to be too drastic. For Spanish gender classification, we started with 67.1% but dropped to 58.9%. Finally for Dutch, we started with 71.9% and dropped to 56.8%.

TABLE V
AGE CLASSIFICATION RESULTS FOR DUTCH

	poly			rbf		
	degree			gamma		
C	1	2	3	0.01	1	100
0.01	0.547	0.513	0.513	0.516	0.589	0.654
1	0.542	0.641	0.649	0.516	0.644	0.717
100	0.539	0.719	0.685	0.646	**0.719**	0.658

It should also be noted that the parameters used in the submitted system differs a bit from the system given here. The system submitted has English to use a radial basis function with gamma and C to be 100. For Dutch and Spanish, the kernel is also a radial basis function with gamma to be equal to 1 and C to be 100. The reason for this difference is that the initial results from previous runs gave these values.

V. CONCLUSION

The conclusion goes here.

VI. RECOMMENDATION

ACKNOWLEDGMENT

The authors would like to thank Erasmus Mundus Mobility for Asia.