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 Gonçalves
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 in forums and chat rooms by getting their linguistic profiles and identifying characteristics and matching it with profiles from known criminals. 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. Another diffuculty in profiling is that text to profiles could be known in one genre and not in the other. For example, it's possible that age and gender could be given in Twitter but not in Blogger. This work aims to explore this problem.

This work investigates author profiling specifically on Twitter data. We first study the accuracy on age and gender classification which are trained and tested on Twitter data. The second part of the study tests models trained on Twitter data on other genres. We used word embeddings as features and evaluated on PAN 2015 and 2016 [16] datasets using the TIRA platform [4], [12].

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 [3] 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 [19] 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 [15] 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 [7] 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. citemeina2013ensemble used collocations and placed first for English while the work of Santosh et al. in [18] worked well with Spanish using POS features.

In PAN 2014 [14], 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 [6], 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 [7]. In [8], 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 [21], 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 [20]. For the work of Marquardt et al. in [9], 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 Naïve Bayes, liblinear, random forests, Support Vector Machines, and decision tables. The method of Lopez-Monroy in [6] gave the best result with an average accuracy of 28.95% on all corpus-types and languages.

In PAN 2015 [13], 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. [2] 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. [5] gave the better result for Italian. This used stylistic features represented by character n-grams and POS n-grams.

Since the problem 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 [10] [11]. 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 task involves two parts. First we want to test the author profiling accuracy of models trained on twitter data against texts from the same genre. The second part is to see if profiling performs well on other genres. This means testing the models trained on twitter against texts on another genre. This is performed on PAN datasets for 2015 and 2016. The interesting thing for PAN is that

A. Datasets

1) PAN 2015 Dataset:

2) PAN 2016 Dataset:

B. 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 [10] and [11]. Essentially, words in a dictionary by a given corpus are initially represented with a vector of random numbers. A word's vector

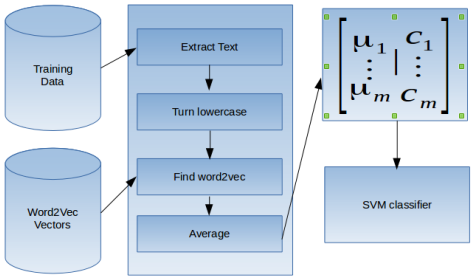

Fig. 1. Overview of the system

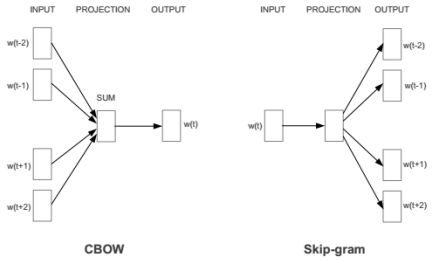

Fig. 2. Diagram for word2vec implementations

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 2. 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.

For our problem, we used wikipedia dumps as an input to the word2vec implementation of gensim [17]. 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 [1].

C. Training and Evaluation

After obtaining word2vec representations for each word as illustrated in figure 3, 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

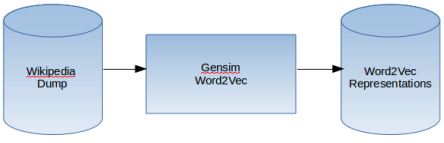

Fig. 3. Overview of word2vec flow.

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

TABLE II
Gender Classification Results for English

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

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 III
Age Classification Results for Spanish

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

TABLE IV
Gender Classification Results for Spanish

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

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 V
Age Classification Results for Dutch

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

Finally, we also add the last table VI 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%.

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.

TABLE VI
PAN RESULTS

	Age	Gender	Joint
English	0.3590	0.6282	0.2179
Spanish	0.4821	0.5893	0.3036
Dutch	0.5680	-	-

V. CONCLUSION

The conclusion goes here.

VI. RECOMMENDATION

ACKNOWLEDGMENT

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

REFERENCES

[1] Training word2vec model on english wikipedia by gensim. http://textminingonline.com/training-word2vec-model-on-english-wikipedia-by-gensim. Accessed: 2010-05-23.

[2] Miguel A Álvarez-Carmona, A Pastor López-Monroy, Manuel Montes-y Gómez, Luis Villaseñor-Pineda, and Hugo Jair-Escalante. Inaoe's participation at pan'15: Author profiling task.

[3] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.

[4] Tim Gollub, Benno Stein, Steven Burrows, and Dennis Hoppe. TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In A Min Tjoa, Stephen Liddle, Klaus-Dieter Schewe, and Xiaofang Zhou, editors, *9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA*, pages 151–155, Los Alamitos, California, September 2012. IEEE.

[5] Carlos E González-Gallardo, Azucena Montes, Gerardo Sierra, J Antonio Núñez-Juárez, Adolfo Jonathan Salinas-López, and Juan Ek. Tweets classification using corpus dependent tags, character and pos n-grams.

[6] A Pastor López-Monroy, Manuel Montes-y Gómez, Hugo Jair Escalante, and Luis Villaseñor-Pineda. Using intra-profile information for author profiling.

[7] Adrian Pastor Lopez-Monroy, Manuel Montes-y Gomez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Esaú Villatoro-Tello. Inaoe's participation at pan'13: Author profiling task. In *CLEF 2013 Evaluation Labs and Workshop*, 2013.

[8] Suraj Maharjan, Prasha Shrestha, and Thamar Solorio. A simple approach to author profiling in mapreduce.

[9] James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. Age and gender identification in social media. *Proceedings of CLEF 2014 Evaluation Labs*, 2014.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[12] Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299, Berlin Heidelberg New York, September 2014. Springer.

[13] Francisco Rangel, P Rosso, M Potthast, B Stein, and W Daelemans. Overview of the 3rd author profiling task at pan 2015. In *CLEF*, 2015.

[14] Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd author profiling task at pan 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes)*, 2014.

[15] Francisco Rangel, Paolo Rosso, Moshe Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 352–365. CELCT, 2013.

[16] Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. Overview of the 4th Author Profiling Task at PAN 2016: Cross-genre Evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2016.

[17] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[18] K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma. Author profiling: Predicting age and gender from blogs. *Notebook Papers of CLEF*, 2013.

[19] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.

[20] Julio Villena-Román and José Carlos González-Cristóbal. Daedalus at pan 2014: Guessing tweet author's gender and age.

[21] Edson RD Weren, Viviane P Moreira, and José PM de Oliveira. Exploring information retrieval features for author profiling—notebook for pan at clef 2014. *Cappellato et al.[6]*.