Author Profiling using SVMs and Word Embedding Averages

Notebook for PAN at CLEF 2016

Roy Bayot and Teresa Gonçalves

Departamento de Informática, Escola de Ciências e Tecnologia,
Universidade de Évora, Rua Romão Ramalho, 59, 7000-671 Évora, Portugal
d11668@alunos.uevora.pt, tcg@uevora.pt

Abstract In this paper, we describe one of the approaches of the participation of Universidade de Évora. Our approach is similar to usual methods where text is preprocessed, features are extracted, and then used in SVMs with cross validation. The main difference is that features used come from averages of word embeddings, specifically word2vec vectors. Using PAN 2016 dataset, we were able to achieve 44.8% and 68.2% for English age and gender classification respectively. We were also able to achieve 51.3% and 67.1% accuracy for Spanish age and gender classification. Finally, we report 71.9% accuracy for Dutch age classification.

1 Introduction

The specific problem associated with PAN 2016's task of author profiling involves the use of training data from a specific corpus and evaluated on a different corpus. The profiling was done in two different dimensions - age and gender classification. The training sets are also available in three different languages - English, Spanish, and Dutch. This is given in full detail in [17]. As with the previous editions of PAN, evaluation is done through the TIRA software as described in [4], and [13].

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 [20] wherein age and gender are related to a specific genre which are blogs, through writing styles. 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 [16] in 2013, the genre focused on was blogs. The task was age and gender profiling for English and Spanish. 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. [10] used collocations and placed first for English while the work of Santosh et al. in [19] worked well with Spanish using POS features.

In PAN 2014 [15], the task was profiling authors with text from four different genres - 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 [22], 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 [21]. 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 [14], 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 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 [11] [12]. 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.

2 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.

2.1 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 [11] and [12]. 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.

For our problem, we used wikipedia dumps as an input to the word2vec implementation of gensim [18]. 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].

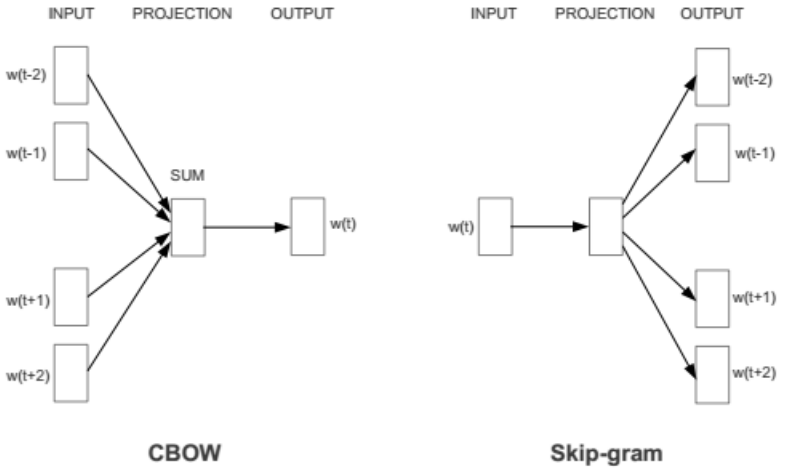

Figure 1. Diagram for word2vec implementations

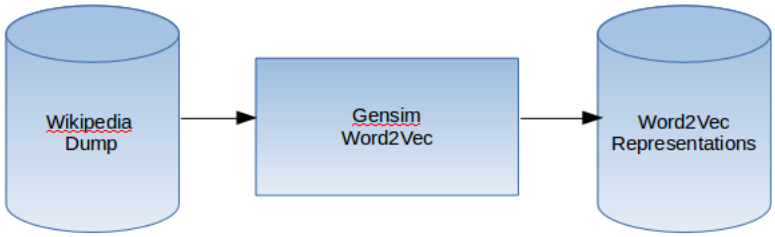

Figure 2. Overview of word2vec flow.

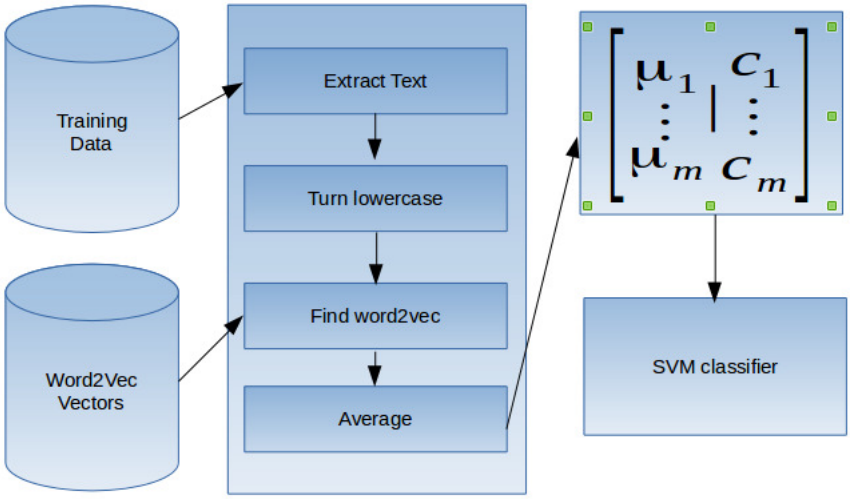

Figure 3. Overview of the system

2.2 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.

3 Results and Discussion

The tables 1- 5 give the all the results for English, Spanish, and Dutch on age and gender using cross validation. Looking at table 1 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 1. Age Classification Results for English using cross validation

	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 3, 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 4, the highest accuracy for gender

Table 2. Gender Classification Results for English using cross validation

	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 3. Age Classification Results for Spanish using cross validation

	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

Table 4. Gender Classification Results for Spanish using cross validation

	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

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.

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 5.

Table 5. Age Classification Results for Dutch using cross validation

	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

Finally, we also add the last table 6 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%.

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 6. PAN 2016 Results

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

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.

4 Conclusion and Recommendations

The use of word embeddings has some merits since the operation is in terms of vector representation and it could be a richer representation. We were able to use this approach

to the current domain of twitter text with modest results. The highest accuracy for age and gender classification for English is 44.8% and 68.2%. For Spanish, age classification yielded 51.3% while gender classification gave 67.1%.

The interesting thing however is that these results came from something simple as not fully preprocessing the input text, as well as using averages of word vectors, and just discarding words that are not in the dictionary, and then just using a Support Vector Machine. The dimensions used were also modest, a mere 100.

There is a lot of room for improvement. One, higher dimension representation could also be used. Word2vec representations trained on twitter data could also yield a better result, and further preprocessing that incorporates twitter specific attributes could be done. Using word vectors also opens the possibility of using deep learning methods with recurrent neural networks and convolutional neural networks.

Acknowledgments We would like to thank Erasmus Mundus Mobility for Asia (EMMA) for the scholarship that enabled this work.

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. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Jair-Escalante, H.: Inaoe's participation at pan'15: Author profiling task
3. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM 52(2), 119–123 (2009)
4. Gollub, T., Stein, B., Burrows, S., Hoppe, D.: TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In: Tjoa, A., Liddle, S., Schewe, K.D., Zhou, X. (eds.) 9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA. pp. 151–155. IEEE, Los Alamitos, California (Sep 2012)
5. González-Gallardo, C.E., Montes, A., Sierra, G., Núñez-Juárez, J.A., Salinas-López, A.J., Ek, J.: Tweets classification using corpus dependent tags, character and pos n-grams
6. López-Monroy, A.P., Montes-y Gómez, M., Escalante, H.J., Villaseñor-Pineda, L.: Using intra-profile information for author profiling
7. Lopez-Monroy, A.P., Gomez, M.M.y., Escalante, H.J., Villasenor-Pineda, L., Villatoro-Tello, E.: Inaoe's participation at pan'13: Author profiling task. In: CLEF 2013 Evaluation Labs and Workshop (2013)
8. Maharjan, S., Shrestha, P., Solorio, T.: A simple approach to author profiling in mapreduce
9. Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M.F., Davalos, S., Teredesai, A., De Cock, M.: Age and gender identification in social media. Proceedings of CLEF 2014 Evaluation Labs (2014)
10. Meina, M., Brodzinska, K., Celmer, B., Czoków, M., Patera, M., Pezacki, J., Wilk, M.: Ensemble-based classification for author profiling using various features. Notebook Papers of CLEF (2013)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)

13. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
14. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: CLEF (2015)
15. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at pan 2014. In: Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes) (2014)
16. Rangel, F., Rosso, P., Moshe Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. pp. 352–365. CELCT (2013)
17. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: 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 (Sep 2016)
18. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
19. Santosh, K., Bansal, R., Shekhar, M., Varma, V.: Author profiling: Predicting age and gender from blogs. Notebook Papers of CLEF (2013)
20. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. vol. 6, pp. 199–205 (2006)
21. Villena-Román, J., González-Cristóbal, J.C.: Daedalus at pan 2014: Guessing tweet author's gender and age
22. Weren, E.R., Moreira, V.P., de Oliveira, J.P.: Exploring information retrieval features for author profiling—notebook for pan at clef 2014. Cappellato et al.[6]