Using TF-IDF n-gram and Word Embedding Cluster Ensembles for Author Profiling

Notebook for PAN at CLEF 2017

Adam Poulston, Zeerak Waseem, and Mark Stevenson

Department of Computer Science
University of Sheffield, UK
{arspoulston1, z.w.butt, mark.stevenson}@sheffield.ac.uk

Abstract This paper presents our approach and results for the 2017 PAN Author Profiling Shared Task. Language-specific corpora were provided for four langauges: Spanish, English, Portuguese, and Arabic. Each corpus consisted of tweets authored by a number of Twitter users labeled with their gender and the specific variant of their language which was used in the documents (e.g. Brazilian or European Portuguese). The task was to develop a system to infer the same attributes for unseen Twitter users. Our system employs an ensemble of two probabilistic classifiers: a Logistic regression classifier trained on TF-IDF transformed n–grams and a Gaussian Process classifier trained on word embedding clusters derived for an additional, external corpus of tweets.

1 Introduction

Author profiling is the task of determining the characteristics of the individual who wrote a document. Many different characteristics can be determined (e.g. personal characteristics such as gender, age, personality [19] and socioeconomic indicators [5,13,14,15]) across a variety of media (e.g. written essays, books, blogs and other social media). Despite their potential ethical concerns, author profiling techniques can be a valuable component in various applications, such as bias reduction in predictive models [2] and language-variant adaption in part-of-speech taggers [1].

In this paper, we present our approach to the 2017 edition of the PAN Author Profiling shared task [10,11,16]. A dataset was provided consisting of Twitter users across four languages and their variants. Each user was labeled with a binary gender label (male/female) and the particular variant of their language (e.g. Brazilian vs European Portuguese). The dataset was balanced by both gender and language variant. Given an unseen user (and their native language), the task is to determine their gender and language variant being used.

To predict gender and language variant, we applied an ensemble of probabilistic machine learning classifiers (described in detail in Section 2). First, an external Twitter corpus was acquired and Tweets geo-located within the countries covered in the tasks languages were extracted (except for the Arabic language variants). This corpus was divided into individual languages (Portuguese, English and Spanish) and used to derive Word2Vec word embeddings [7,8] for each language. Then, each set of language

specific word embeddings were clustered using K-Means to derive a set of word to cluster mappings, which can be thought of as roughly analogous to topics in a topic model. The normalised frequency of each word cluster across a user's tweets was used to train a Gaussian Process classifier. Second, a Logistic Regression classifier was then trained using TF-IDF transformed unigram and bigram frequencies. Both classifiers were employed in an ensemble approach by averaging the predicted probabilities for each sample to determine the label.

2 Approach

Our approach combines two probabilistic classifiers trained on distinct feature sets in an ensemble to predict gender and language variant. Two classifiers were applied: a Logistic Regression classifier trained on TF-IDF n–grams (Section 2.1) and a Gaussian Process classifier trained on word cluster frequencies (Section 2.2). For each unseen document, probabilities from both classifiers are taken and averaged, and the highest average probability class is taken as the prediction. Models were trained using the implementations found in scikit-learn [9] unless stated otherwise.

For Arabic data, only the Logistic Regression classifier is applied, as the volume of geo-located Arabic tweets collected was too low to allow for training of robust Word2Vec models for use with the Gaussian Process classifier.

2.1 Logistic regression classifier with TF-IDF n–grams

Word unigram and bigram features were extracted for each training document. The text was tokenised using a Twitter-aware tokeniser [4]; no additional steps were taken to deal with the extra complexities of Arabic text. A list of stop words was not used while deriving n–gram features, instead tokens that appeared in more than 90% of the documents were removed, as this allows for the removal of n–grams common across a language's variants while also removing stop words.

TF-IDF weighting was applied to down-weight n–grams common across the documents and assign a higher weight to n–grams which are rare.

A Logistic Regression classifier was trained for each language using the n–gram features. Logistic Regression was chosen for use with the n–gram features because it has been shown to perform well on similar high-dimensional classification tasks, and produces probabilistic predictions [3].

2.2 Gaussian process classifier with word embedding clusters

We obtained the data for our word embedding clusters from a Twitter Firehose[1] sample collected throughout 2015. We only used tweets that were geo-located in the specific language regions determined by the shared task (see Table 1).

Some language variants were less frequent in the resulting datasets than others, for instance we collected very few tweets from Ireland compared to the U.S.A. Down-sampling was used to avoid over representation of the more prevent language variants.

[1] Twitter Firehose has since been discontinued and can no longer be accessed.

Table 1. Countries scraped for each language.

English (F_{en})	**Spanish** (F_{sp})	**Portuguese** (F_{pt})
Australia	Argentina	Brazil
Canada	Chile	Portugal
Great Britain	Colombia	
Ireland	Mexico	
New Zealand	Peru	
United States	Spain	
	Venezuela	

Data for the language variant with the largest volume of documents was reduced so that it contained no more than 10 times number of tweets of the smallest language variant.

Word embeddings For each language dataset (F_{en}, F_{es}, and F_{pt}) were trained using the Word2Vec [7,8] implementation in gensim [18] with Continuous Bag of Words (CBOW), negative sampling, 200 dimensions, and a window size of 10.

We applied K-Means clustering [6] to the word embeddings to derive a set of 100 clusters for each language, in which each word is assigned a cluster based on its nearest cluster in the embedding space. We then computed the frequency distribution of the clusters for every training document, and used them as features to train a Gaussian Process classifier with an RBF kernel [17].

Similar word embedding clusters have been applied with Gaussian Processes to perform other author profiling tasks such as socio-economic status detection [5]; furthermore, the derived clusters are similar to topics derived in a topic model, in that they identify semantically similar groups of words in documents, which we found to perform well in a similar task [12].

3 Results

Table 2 shows the accuracy scores achieved by a Support Vector Machine (SVM) classifier with a linear kernel, trained on the same TF-IDF n–grams described in Section 2.1. We chose this approach as our baseline, as it has been shown to perform well on similar tasks and represent a strong baseline.

Table 2. Baseline accuracy scores for gender and language variant prediction for each language derived from a SVM classifier trained on TF-IDF n–grams.

Target	**Spanish**	**English**	**Portuguese**	**Arabic**
Gender	0.7361	0.7896	0.8263	0.7450
Language variant	0.9532	0.8617	0.9800	0.8150
Joint	0.7007	0.6838	0.8113	0.6275

Table 3. Accuracy scores for gender and language variant prediction for each language as submitted for the PAN: Author Profiling task 2017.

Target	Spanish	English	Portuguese	Arabic
Gender	0.7939	0.7829	0.8388	0.7738
Language variant	0.9368	0.8038	0.9763	0.7975
Joint	0.7471	0.6254	0.8188	0.6356

Table 3 shows the results of our final submitted run for the PAN: Author Profiling task 2017. For Spanish, English and Portuguese the results were attained by applying the ensemble of Logistic Regression and Gaussian Process classifiers described in Section 2; for Arabic only the Logistic regression classifier was applied (Section 2.1). In the rankings for the PAN Author Profiling shared task [16], our approach achieved 7th place out of 22 entries for joint prediction and 6th for gender, exceeding reported baselines. We achieved poorer results for language variant prediction at 9th place, and did not exceed the baseline approach.

3.1 Discussion

In Table 3, we see that the our ensemble performs quite well for identifying language variant or gender individually. For joint prediction our ensemble performs less well, likely due to errors in either gender or language variant prediction propagating through to incorrect joint predictions. Of the three languages the ensemble was applied to, the best performance was observed for Portuguese and the worst for English. Broad topics of interest appear to be effective for the gender prediction problem while individual terms that are unique to specific language variants are more discriminating for language variant prediction.

Similar to our results in a previous PAN: Author Profiling Profiling shared task entry [12], in which LDA topic models were able to improve predictive performance over word n–grams, word embedding clusters improved predictive accuracy for gender classification. For the language variant differentiation task, introducing the word embedding clusters in fact reduced accuracy scores over earlier runs.

Under our current clustering scheme, each term was assumed to be equally as representative of its cluster as each other term; in practise though, certain terms were closer to the centroid in embedding space than others. Prior to submission we had begun experimenting with weighting terms based on their proximity to their closest centroid, and our initial findings were promising. In future work we would like to investigate the effect of weighting terms in more detail.

4 Conclusion

In this notebook, we have shown that by employing an ensemble of classifiers and utilising clusters of word embeddings reasonable results can be achieved. We propose,

that our approach can be improved by weighting the word embedding clusters by the distance to the cluster centroid.

References

1. Blodgett, S.L., Green, L., O'Connor, B.: Demographic dialectal variation in social media: A case study of african-american english pp. 1119–1130 (November 2016)
2. Culotta, A.: Reducing sampling bias in social media data for county health inference. In: Joint Statistical Meetings Proceedings (2014)
3. Freedman, D.A.: Statistical models: theory and practice. cambridge university press (2009)
4. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.a.: Part-of-speech tagging for Twitter: annotation, features, and experiments. Human Language Technologies 2(2), 42–47 (2011)
5. Lampos, V., Aletras, N., Geyti, J.K., Zou, B., Cox, I.J.: Inferring the socioeconomic status of social media users based on behaviour and language (2016)
6. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297. Oakland, CA, USA. (1967)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
8. 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)
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
10. 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)
11. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Initiative (CLEF 17). Springer, Berlin Heidelberg New York (Sep 2017)
12. Poulston, A., Stevenson, M., Bontcheva, K.: Topic models and n-gram language models for author profiling-notebook for pan at clef 2015. (2015)
13. Poulston, A., Stevenson, M., Bontcheva, K.: User profiling with geo-located posts and demographic data pp. 43–48 (November 2016)
14. Preoţiuc-Pietro, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through Twitter content. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) pp. 1754–1764 (2015)
15. Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., Aletras, N.: Studying user income through language, behaviour and affect in social media. PloS one 10(9), e0138717 (2015)

16. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (sep 2017)
17. Rasmussen, C.E., Williams, C.K.: Gaussian processes for machine learning, vol. 1. MIT press Cambridge (2006)
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)
19. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P., Ungar, L.H.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS ONE 8(9), e73791 (09 2013)