Topic Models and n–gram Language Models for Author Profiling

Notebook for PAN at CLEF 2015

Adam Poulston, Mark Stevenson, and Kalina Bontcheva

University of Sheffield, UK
{arspoulston1,mark.stevenson,k.bontcheva}@sheffield.ac.uk

Abstract Author profiling is the task of determining the attributes for a set of authors. This paper presents the design, approach, and results of our submission to the PAN 2015 Author Profiling Shared Task. Four corpora, each in a different language, were provided. Each corpus consisted of collections of tweets for a number of Twitter users whose gender, age and personality scores are know. The task was to construct some system capable of inferring the same attributes on as yet unseen authors. Our system utilizes two sets of text based features, n–grams and topic models, in conjunction with Support Vector Machines to predict gender, age and personality scores. We ran our system on each dataset and received results indicating that n-grams and topic models are effective features across a number of languages.

1 Introduction

Author profiling is the problem of determining the characteristics of a set of authors based on the text they produce, how they behave and with whom they interact. An author profiling task will typically centre on predicting one or more *attributes* of one or more *authors*. An attribute can represent any element of a persons self, ranging from obvious outward characteristics such as gender and age, to more personal qualities such as personality, political leaning or sexual orientation [5, 1, 14, 12, 6].

A range of potential applications exist for author profiling techniques, many of which give rise to deep ethical considerations. A company or organisation could use an author profiling tool to identify their core user-base. Marketers could further target advertisement to social media users who are determined to hold particular characteristics. Law enforcement could potentially use such a system to link on-line criminal behaviour with individuals. Studies have already investigated the use of author profiling techniques in identifying on-line grooming [9].

A machine learning approach was employed for this task in order to predict gender, age and personality. Topic models, implemented using Latent Dirichlet Allocation (LDA) [2], and n–gram language models were used to extract features to train Support Vector Machine (SVM) classifiers (for gender and age) and regressors (for personality dimensions).

2 Task Outline

For the Author Profiling task at PAN 2015, a set of Twitter users whose gender, age and personality is known is provided. These users are further divided into four languages: Italian, English, Dutch and Spanish. The task is, given a single set of these users, some judgement of age, gender and personality must be made on as yet unseen users [11].

Four corpora of tweets of different languages are provided. The corpora are balanced by author gender, such that there is an equal number of male and female authors present in each corpus. There is no guarantee that each author has the same number of tweets, and as such over-fitting to particular authors is a risk. For age there is definite imbalance, with particular age groups containing many more authors.

The task of determining age in this case has been converted to a classification problem, where a range of ages is to be predicted rather than a continuous value. Gender is also a classification problem; binary selection of male or female.

Personality prediction in this task was to estimate each user's "Big 5" personality scores, in the range of -0.5 to 0.5, and is treated as a regression problem. The personality dimensions considered are all of the Big 5: openness, conscientiousness, extraversion, agreeableness, and neuroticism.

3 Data and Preprocessing

The main pre-processing step undertaken was tokenisation using a Twitter specific tokeniser [4].

In early experiments on the data, all short-links present in the text were followed and converted to the domain name of the website found, as previous author profiling studies have identified website use as a potential analogue for some attributes [7, 3]. This was discarded in the final approach as no improvement could be noted with its inclusion. A similar experiment was also performed to replace all links with a single "link present" token, but again no improvement was noted.

The Twitter specific step of eliminating "retweets" was also considered, although the provided data contains so few retweets this step was deemed unnecessary. In most other Twitter profiling tasks this would be included. Another consideration is that some Tweets are in the form "shared via some app", and do not register as retweets. These are not considered in the scope of this shared task, but may be a useful addition in future experiments.

4 Feature Extraction

In the final approach word n–grams and topics from topic models were used as features. Other features were experimented with in early development, but discarded due to poor performance. In this section the features experimented with are presented and discussed. In order to assess the affect of various features a 10-fold cross validation was performed on the training data.

n–gram language model Throughout early experiments it became apparent that unigrams and bigrams together produced the most reliable results and as such would form the basis of any system developed. n–grams were weighted using the tf-idf term weighting scheme, where a term's rating is based not only on its frequency in a document, but also against how common the term is in the whole set of documents, rating very common terms lowly and uncommon terms highly.

A stop-list was not used in building the n–gram feature vectors due to the multilingual nature of the problem, instead all tokens that appeared in more than 70% of the documents, as this is a roughly analogous, language independent technique.

Topic model Topic models are a group of algorithms that identify hidden themes (topics) in collections of documents. The topic model used in this approach is Latent Dirichlet Allocation [2], a generative model in which documents are modelled as a finite mixture of topics, such that each word in a document must be generated by one of its topics. Topic models were implemented using the library gensim [13]. Topic models have been shown to produce reliable results when used alone and in conjunction with other features [10, 15].

As part of the training process an LDA topic model is trained on the input data, with a target of 10 topics. Ideally the model would be trained on a large additional corpus to produce more robust topics, sadly due to time and computational constraints this was not possible in the scope of this shared task.

The trained model is then used to infer topics, labelled as present or not, on unseen documents. There is also the option to weight a topic feature by the likelihood that it belongs to the input text, although early experiments showed that this added no benefit.

Parts–of–speech In early experiments all tweets were POS tagged as part of the pre–processing step using a Twitter specific part–of–speech tagger [4]. Various studies have identified POS tags as a useful feature [12, 15], and despite some improvement being noted, they were not included as a feature in the final submission, as the part–of–speech tagger used was English specific, and as such would not be compatible with the other three languages. In future it would be interesting to examine their affect on non-English results.

5 Assessing Features

As official results were not consistently available throughout development, 10-fold cross validations was used throughout development, to assess the affect of different features on classifier accuracy. Results from this cross-validation, which motivated feature choice in the final submission, are presented in Table 2. The feature(s) with the best score for each attribute for each language is highlighted in bold.

Results are presented in each language for n–gram features, LDA features, and the two in conjunction. In the English case, results for POS tagged n–grams are also included. These results show POS tagged n–grams as being the best feature for English gender and age prediction; despite this they were not used in the final submission, as a comparable POS tagger could not be found for Spanish, Dutch and Italian tweets.

In most cases n–gram features provided the best results, but not by a significant margin, with n–grams in conjunction with LDA topics performing similarly. LDA topics on their own proved to be a very poor quality for the English and Spanish datasets, and gave the worst results in all cases.

The final submission included n–grams in conjunction with LDA topics, as these judgements proved to be more stable across folds than n–grams on their own.

		Accuracy		**Root Mean Squared Error**				
Language	**Features**	**Gender**	**Age**	**E**	**N**	**A**	**C**	**O**
English	n-gram	**0.7754**	0.7245	**0.1510**	**0.1876**	0.1568	**0.1410**	**0.1281**
	LDA	0.5062	0.4683	0.1949	0.2424	0.1776	0.1686	0.1625
	n-gram + LDA	0.7500	**0.7438**	0.1559	0.2010	**0.1522**	0.1422	0.1327
	POS	*0.7758*	*0.7829*	*0.1561*	*0.2026*	*0.1700*	*0.1443*	*0.1348*
Spanish	n-gram	**0.8800**	**0.7300**	**0.1501**	**0.1691**	**0.1426**	**0.1468**	**0.1520**
	LDA	0.5400	0.4100	0.1715	0.2469	0.1795	0.2199	0.1967
	n-gram + LDA	0.8000	0.7200	0.1537	0.1831	0.1502	0.1617	0.1550
Dutch	n-gram	**0.8250**	N/A	**0.1112**	**0.1754**	**0.1374**	**0.1039**	**0.1123**
	LDA	0.7083	N/A	0.1618	0.2366	0.1873	0.1355	0.1470
	n-gram + LDA	0.7083	N/A	0.1307	0.1845	0.1476	0.1162	0.1165
Italian	n-gram	**0.8500**	N/A	**0.1208**	**0.1600**	**0.1283**	**0.1110**	**0.1377**
	LDA	0.6000	N/A	0.1963	0.2602	0.2150	0.1565	0.2441
	n-gram + LDA	0.7083	N/A	0.1461	0.1670	0.1492	0.1190	0.1442

Table 1. Classifier accuracy and mean squared error results from cross validation on training data

6 System Architecture

The architecture of the submitted system is presented in Figure 1. The system comprises two main components: a model generation module, and one which uses a pre-trained model to infer the attributes it contains on unseen documents.

For model generation the training data is fed through several feature extraction modules. Firstly, an LDA model is trained which is then used in the "Topic Extraction" module. The same data is also passed through an "n–gram Extraction" module. The resulting feature vectors are then used to train a machine learning model.

The machine learning algorithm used in the final submission is Support Vector Machines (SVM) as they have been repeatedly shown to produce better results than other algorithms. Experiments were performed with ensemble methods and other algorithms, but none beat the results achieved by the SVM implementation.

For age and gender a Support Vector Classifier with a linear kernel was used. For the personality recognition element Support Vector Regressors were used, again with a linear kernel. All implementations were provided in Scikit-learn [8].

The resulting model can then be presented with previous unseen documents, and perform judgements on the author attributes it was trained with.

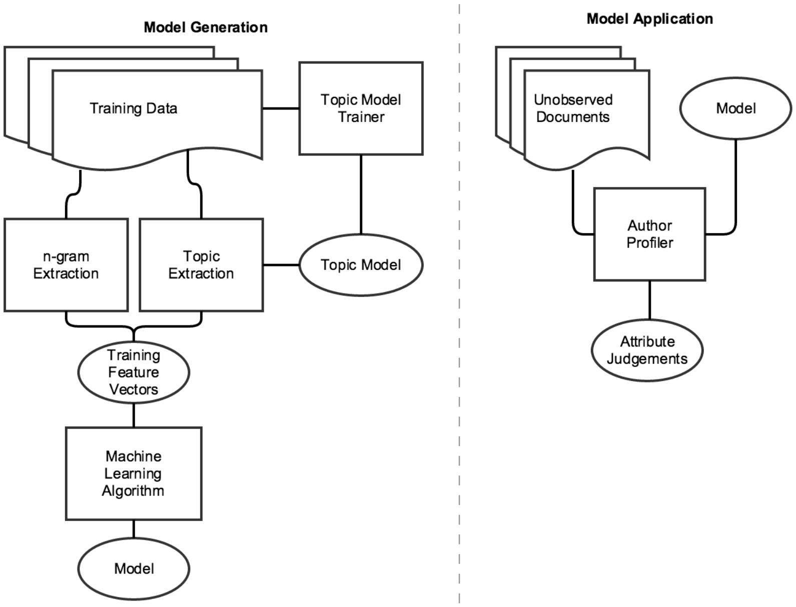

Figure 1. Architecture of presented system.

7 Results

The results of the final run are presented in Table 2. The system performed best on the Italian dataset, achieving a global score above 0.8, where scores for submitted systems ranged from 0.8658 to 0.6024. For the English and Spanish corpora scores were in the ranges 0.7906 to 0.5217 and 0.8215 to 0.5049 respectively, with the results obtained by our system falling roughly in the middle of these ranges. The worst performance was obtained for the Dutch dataset, scoring on the bottom end of the range 0.9406 to 0.6703.

In most cases the final results are worse than those observed by applying cross-validation to the training data. However similar or better results were observed for some personality elements across languages. English age prediction and Spanish gender prediction also achieved reasonable scores compared to the cross-validation.

The results show that n–grams and topic models are a useful element in developing author profiling systems across a number of languages and provide reasonable results without any additional features. In order to improve the system without adding any other features the LDA topic model could be trained on a large external corpus of text, in theory leading to a more robust model. Additional stylometric features such as readability and text structure could also be applied to assess their affect on performance. It would also be interesting to asses the effect of network and behavioural features on performance should an additional dataset containing appropriate information become available.

			Accuracy			**Root Mean Squared Error**				
Language	**Global**	**RMSE**	**Gender**	**Age**	**Joint**	**E**	**N**	**A**	**C**	**O**
English	0.6743	0.1725	0.6901	0.7394	0.5211	0.1381	0.2223	0.1918	0.1749	0.1352
Spanish	0.6918	0.1619	0.8409	0.5909	0.5455	0.1669	0.2285	0.1398	0.1412	0.1329
Italian	0.8061	0.1378	0.7500	N/A	N/A	0.1279	0.1923	0.1257	0.1187	0.1243
Dutch	0.6796	0.1409	0.5000	N/A	N/A	0.1752	0.1511	0.1444	0.1344	0.0993

Table 2. Results of final software submission including global rankings and individual attribute performance

8 Conclusion

In this document we have presented our approach to the PAN 2015 Author Profiling shared task. We used Support Vector Machine classifiers and regressors in conjunction with n–gram and topic features, in order to provide judgements on age, gender and personality.

In future work we would like to investigate the effect of additional text and non-text features on classifier performance, as well as an investigation into system performance on larger datasets.

Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-14-1-0333.

References

[1] Argamon, S., Dhawle, S., Koppel, M., Pennebaker, J.W.: Lexical Predictors of Personality Type. In: In Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America (2005)

[2] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3(4-5), 993–1022 (2012)

[3] Conover, M.D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of twitter users. In: Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011. pp. 192–199 (2011)

[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] Koppel, M.: Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing 17(4), 401–412 (2002)

[6] Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences of the United States of America 110(15), 5802–5 (2013)

[7] Michelson, M., Macskassy, S.A.: What blogs tell us about websites: a demographics study. In: Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11. pp. 365–374 (2011)

[8] 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 {P}ython. Journal of Machine Learning Research 12, 2825–2830 (2011)

[9] Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting age and gender in online social networks. In: International Conference on Information and Knowledge Management, Proceedings. pp. 37–44 (2011)

[10] Pennacchiotti, M., Popescu, A.M.: A Machine Learning Approach to Twitter User Classification. In: ICWSM. pp. 281–288 (2011)

[11] Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) CLEF 2015 Labs and Workshops, Notebook Papers. vol. 1391. CEUR-WS.org (2015)

[12] Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. Proceedings of the 2nd international workshop on Search and mining user-generated contents - SMUC '10 p. 37 (2010)

[13] Rehurek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora pp. 45–50 (May 2010)

[14] Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. In: Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs: Papers from the AAAI Spring Symposium. pp. 199–205 (2006)

[15] 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 (Jan 2013)