AFRL-AFOSR-UK-TR-2016-0011

Individual Profiling using Text Analysis 140333

Mark Stevenson
UNIVERSITY OF SHEFFIELD, DEPARTMENT OF PSYCHOLOGY

04/15/2016
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOE
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
26-04-2016	Final	15 Sep 2014 to 14 Sep 2015

4. TITLE AND SUBTITLE
Individual Profiling using Text Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-14-1-0333

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Mark Stevenson

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF SHEFFIELD, DEPARTMENT OF PSYCHOLOGY
FIRTH CT
SHEFFIELD, S10 2TP GB

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
EOARD
Unit 4515
APO AE 09421-4515

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOE

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-UK-TR-2016-0011

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Author profiling is the task of determining the attributes for a set of authors. This report presents the design, approach, and results of our approach to using data from the PAN 2015 Author Profiling Shared Task to predict personal attributes, as per the project brief. 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 known. 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. Thes

15. SUBJECT TERMS
Computer Science, Text Analysis, Latent Dirichlet Allocation, EOARD

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			JORDAN, JEREMY
Unclassified	Unclassified	Unclassified	SAR		19b. TELEPHONE NUMBER *(Include area code)* 011-44-1895-616002

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Research Title:
Individual Profiling Using Text Analysis

Grant Number:
FA9550-14-1-0333

PI Name:
Dr. Mark Stevenson, University of Sheffield

Period of Performance:
15 September 2014 to 14 September 2015

Contents

1 Summary **1**

2 Introduction **1**

2.1 Author Profiling 1

2.2 Task Outline 2

3 Methods, Assumptions and Procedures **3**

3.1 Data and Preprocessing 3

3.2 Feature Extraction 3

3.3 Assessing Features 5

3.4 System Architecture 5

4 Results and Discussion **6**

5 Conclusions **8**

1 Summary

Author profiling is the task of determining the attributes for a set of authors. This report presents the design, approach, and results of our approach to using data from the PAN 2015 Author Profiling Shared Task to predict personal attributes, as per the project brief. 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. These results have also been submitted to PAN at CLEF 2015 [22].

2 Introduction

2.1 Author Profiling

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 [10, 2, 28, 26, 12].

Techniques have been shown to be viable across a wide range of attributes and domains. Earlier studies focussed on traditional media such as the British National Corpus [10] and student essays, as well as digital yet still formal media such as email [5] and transcriptions of interview speech [16, 14]. With the advent of open online platforms, many later studies took on a digital approach, focussing on more casual platforms such as blogs [28] and social media [20]. Despite this studies in traditional media continue to be relevant, for example, one study attempts to identify political bias in supposedly impartial texts [9].

Online platforms provide authors with options that do not exist in traditional media. Blogs and other internet media, for example, often provide users with mechanisms to alter the formatting of their text either with simple menu options or a more technical approach with HTML, CSS or some

custom styling language. It is possible that these formatting choices could be indicitive of some attributes [28], with one aproach [1] predicting author gender fairly accurately with only profile background color!

As well as covering many domains, a wide variety of attributes have been tackled with varying degrees of succes. Early studies focussed on more obvious traits such as age and gender, beggining with [10] study of gender in the British National Corpus. Additional studies soon followed, looking at many additional aspects such as personality [2], first language [11] and level of education [6, 7].

With the advent of social media and blogs, data became much more readily available, as such many studies were repeated and extended to cover social media, as well as investigations into new attributes such as ethnicity [25], political ideology [26] and sexual preference [12].

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 [20].

The project brief specified that the affect of including LDA topics in with traditional text features was to be examined. As such a machine learning approach was employed to predict gender, age and personality. Topic models, implemented using Latent Dirichlet Allocation (LDA) [3], 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).

2.2 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 [24].

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 Methods, Assumptions and Procedures

3.1 Data and Preprocessing

The text provided proved to be quite clean and little pre-processing was required other than tokenisation. Despite this, other potential avenues were investigated.

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 [18, 4]. 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.

3.2 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 multi-lingual 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 [3], 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 [27]. Topic models have been shown to produce reliable results when used alone and in conjunction with other features [21, 29].

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 [8]. Various studies have identified POS tags as a useful feature [26, 29], 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.

3.3 Assessing Features

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.

3.4 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 [19].

		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

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

4 Results and Discussion

The results of the final system run submitted to PAN 2015 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.

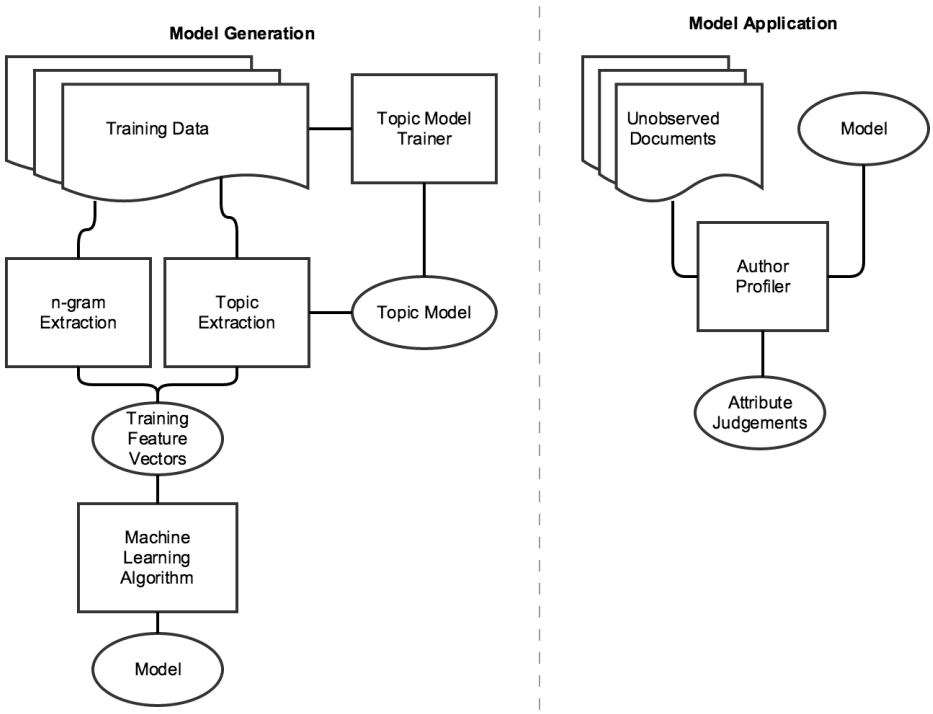

Figure 1: Architecture of presented system.

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.

The way an author behaves in the context of interacting with their medium (be it social media, conversation or essay writing) has, in other studies, be telling of their characteristics. For example, according to the big five model of personality an extroverted person is likely to be more outgoing, assertive and have a positive demeanour [17]. Conversational elements have also been shown to be useful [16, 14, 15].

It is also possible to code for behaviour in online media. Studies have identified varying patterns of social media activity times in areas of high and low unemployment, with those low employment areas seeing a sharp rise in posts around the start of the working day [13]. Other studies have attempted to detect conversational behaviours on social media, as earlier research showed them to be of use for author profiling.

An analysis of an authors social network can also give rise to interesting judgements about them. It has been shown for example, that the presence of certain "Likes" made by an author on the platform Facebook, can be indicative of wide number of characteristics. Other social network properties may also be useful, in [23] four distinct groups of users, where each group has similar personality scores, were identified, based on a user's tendency to follow, be followed and favourite tweets on Twitter.

For the purpose of this project hovever, these techniques were not further investigated, due to the format of the provided data, although in future it would be very interesting to assess their effect on system performance.

			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

5 Conclusions

In this document we have presented our findings regarding the affect of the inclusion LDA topics in conjunction with traditional text features. 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. Our findings indicate that the addition of LDA topics does improve system performance in most cases.

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.

References

[1] Jalal S. Alowibdi, Ugo a. Buy, and Philip Yu. Language independent gender classification on Twitter. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, (May):739–743, 2013.

[2] Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. Lexical Predictors of Personality Type. In *In Proceedings of the 2005 Joint Annual Meeting of the Interface and the Clas- sification Society of North America*, 2005.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2012.

[4] Michael D. Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 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*, pages 192–199, 2011.

[5] OY de Vel, MW Corney, and AM Anderson. Language and gender author cohort analysis of e-mail for computer forensics. *Digital Forensics Research Workshop*, pages 1–16, 2002.

[6] Dominique Estival, Tanja Gaustad, and SB Pham. TAT: an author profiling tool with application to Arabic emails. *Proceedings of the Australiasian Language Technology Workshop*, pages 21–30, 2007.

[7] Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. Author Profiling for English Emails. *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272, 2007.

[8] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah a Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. *Human Language Technologies*, 2(2):42–47, 2011.

[9] Zubin Jelveh, Bruce Kogut, and Suresh Naidu. Detecting Latent Ideology in Expert Text: Evidence From Academic Papers in Economics. *anthology.aclweb.org*, (2013):1804–1809, 2014.

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

[11] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Determining an Author's Native Language by Mining a Text for Errors. *. . . on Knowledge discovery in data mining*, pages 624–628, 2005.

[12] Michal Kosinski, David Stillwell, and Thore Graepel. 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.

[13] Alejandro Llorente, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment. *arXiv preprint arXiv:1411.3140*, pages 1–19, 2014.

[14] François Mairesse and Marilyn a. Walker. Automatic recognition of personality in conversation. *Proceedings of the Human Language Technology Conference of the NAACL*, (June):85–88, 2006.

[15] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.

[16] Francois Mairesse and Marilyn Walker. Words Mark the Nerds : Computational Models of Personality Recognition through Language. *28th Annual Conference of the Cognitive Science Society*, 2006.

[17] Gerald Matthews, Ian J Deary, and Martha C Whiteman. *Personality traits*. Cambridge University Press, 2003.

[18] Matthew Michelson and Sofus A. Macskassy. 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*, pages 365–374, 2011.

[19] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] Claudia Peersman, W Daelemans, and L Van Vaerenbergh. Predicting age and gender in online social networks. In *International Conference on Information and Knowledge Management, Proceedings*, pages 37–44, 2011.

[21] Marco Pennacchiotti and Ana-Maria Popescu. A Machine Learning Approach to Twitter User Classification. In *ICWSM*, pages 281–288, 2011.

[22] Adam Poulston, Mark Stevenson, and Kalina Bontcheva. Topic Models and n–gram Language Models for Author Profiling – Notebook for PAN at CLEF 2015. page To appear, 2015.

[23] Daniele Quercia and Michal Kosinski. Our Twitter Profiles,Our Selves: Predicting Personality with Twitter. 2011.

[24] Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd Author Profiling Task at PAN 2015. In *Working Notes Papers of the CLEF 2015 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2015.

[25] Delip Rao, M. Paul, Clay Fink, D. Yarowsky, Timothy Oates, and G. Coppersmith. Hierarchical Bayesian Models for Latent Attribute Detection in Social Media. In *Fifth International AAAI Conference on Weblogs and Social Media*, pages 598–601, 2011.

[26] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. *Proceedings of the 2nd international workshop on Search and mining user-generated contents - SMUC '10*, page 37, 2010.

[27] Radim Rehurek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. pages 45–50, May 2010.

[28] Jonathan Schler, M Koppel, S Argamon, and J Pennebaker. 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*, pages 199–205, 2006.

[29] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E P Seligman, and Lyle H Ungar. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PloS one*, 8(9):e73791, January 2013.

1.

1. Report Type

Final Report

Primary Contact E-mail
Contact email if there is a problem with the report.

jeremy.jordan@us.af.mil

Primary Contact Phone Number
Contact phone number if there is a problem with the report

+44-1-8956-16002

Organization / Institution name

AFOSR/EOARD

Grant/Contract Title
The full title of the funded effort.

Individual Profiling using Text Analysis

Grant/Contract Number
AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-14-1-0333

Principal Investigator Name
The full name of the principal investigator on the grant or contract.

Mark Stevenson

Program Manager
The AFOSR Program Manager currently assigned to the award

Jeremy Jordan

Reporting Period Start Date

09/15/2014

Reporting Period End Date

09/14/2015

Abstract

Author profiling is the task of determining the attributes for a set of authors. This report presents the design, approach, and results of our approach to using data from the PAN 2015 Author Profiling Shared Task to predict personal attributes, as per the project brief. 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 known. 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. These results have also been submitted to PAN at CLEF 2015.

Distribution Statement
This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement
If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form

Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF The maximum file size for an SF298 is 50MB.

SF 298.pdf

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF . The maximum file size for the Report Document is 50MB.

Final report.pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

PAN 2015 Competition Conference Proceedings

Changes in research objectives (if any):

None

Change in AFOSR Program Manager, if any:

None

Extensions granted or milestones slipped, if any:

None

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

Apr 14, 2016 10:59:57 Success: Email Sent to: jeremy.jordan@us.af.mil