Cross Genre Author Profiling: Predicting Age and Gender based of text of different genres

Nishant Shah
nishshah@indiana.edu
Indiana University

Abstract

Author profiling is used to determine an author's gender, age, native language, personality type, etc. Author profiling is a problem of growing importance in a variety of areas, including forensics, security and marketing. In this paper, I am discussing the gender and age prediction problem of the author profiling problem in English from a cross-genre perspective.

Keywords
Author Profiling, Age prediction, Gender prediction, Natural Language Processing, Machine Learning, Social Media, Text Analysis

1. Introduction and Motivation

When social media and information technology are heavily influencing and impacting our lives, we rarely do have information about the creators of these information. The possibility of determining people's traits on the basis of what they write is a field of growing interest named author profiling. Author profiling distinguishes between classes of authors by studying their sociolect aspect, i.e., how language is shared or how an author can be characterized from a psychological viewpoint. Authorship analysis aims at finding out as much information as possible about a person by analyzing the text written by that person. The profiling focuses on identifying author attributes like gender, age, native language, personality type, level of education etc. Author profiling has wide range of applications in forensics, marketing and internet security. Eg. companies may analyze the texts of product reviews to identify which type of customer likes or dislikes their products, or the police may identify the perpetrator of a crime by analyzing suspects' writing profiles.

In the age of information overload and introduction of powerful computation systems, this area is getting increasing importance from past 3-5 years. As a result, PAN has been organizing various shared tasks on digital forensics from 2013 and author profiling is one of such task. Before PAN also, many independent groups having background of computational linguistics, statistics and computer science have tried to address this task. And they have come up with interesting findings. In this paper, I am leveraging their research and have made an attempt for finding good classification approach for author profiling task. I am also surveying different types of features which I explored but end up not using them because of various reasons explained in appropriate section.

Problem Statement for this paper:

Most of the previous work related to author profiling has been performed on same genre documents. Systems built on such same genre often don't work with text of different genre. I will be focusing on age and gender prediction task of author profiling. My aim for this project is to build such a system which work genre independently. Below is the outline based on which, my profiling system will work.

Organization of remainder of the paper:

The remainder of the paper is organized as follows. Section 2 outlines previous state of the art work. Section 3 describes the various corpus used in this paper, how they are related to the specific task of cross genre author profiling and how I incorporated those data into author profiling task. Section 4 explains different experiments performed. Section 5 depicts the results of experiments described in section 4. Section 6 gives summary and conclusion. Section 7 outlines opportunities and directions for future work. Section 7 lists references.

2. Related Work

The study of how certain linguistic features vary according to the profile of their authors has been a subject of interest from more than a decade before from different areas such as psychology, linguistics and, natural language processing.

Pennebaker et al. researched how the style of writing is associated with personality traits, studying how the variation of linguistic characteristics in a text can provide information regarding the gender and age of its author. Argamon et al. investigated the task of gender identification on the British National Corpus and achieved approximately 80% accuracy. Holmes and Meyerhoff et al., Burger and Henderson et al. have also investigated how to obtain age and gender information from formal texts.

In advent of information technology, most of the recent research is focused on various social media data which is more colloquial, less formal and less structured data like Twitter posts, Blog/Facebook posts etc.

Koppel et al. studied the problem determining an author's gender by combining simple lexical and syntactic features, and achieved approximately 80% accuracy. Schler et al. studied the effect of age and gender in the writing style in blogs; they gathered over 71,000 blogs and obtained a set of stylistic features like non-dictionary words, parts-of-speech, function words and hyperlinks, combined with content features, such as word unigrams with the highest information gain. They obtained an accuracy of 80% for gender identification and 75% for age identification. They found and demonstrated that language features in blogs correlates with age, eg. the use of prepositions and determiners. Goswami et al. added some new features as slang words and the average length of sentences, improving accuracy to 80.3% in age group identification and to 89.2% in gender detection. Zhang and Zhang experimented with short segments of blog post, and obtained 72.1% accuracy for gender prediction.

In Author Profiling shared task at PAN 2013-16, most of the participant used combination of stylistic and content based features. Stylistic features such as frequency of punctuation marks, quotations, capital letters, n-gram POS taggers, emoticons, readability features, HTML information like number of URLs. Content based features included topic modeling using LDA, latent semantic analysis, bag of words, TF-IDF, dictionary based words, second order representations based on relationships between documents and profiles. Some participants used Information Retrieval approaches such as cosine similarity, Okapi BM25.

3. Data and feature Engineering

3.1 Corpuses

I have used 4 corpuses for this paper, they are explained in following subsections. These corpuses are available in more than 1 different language but I am using only English language.

3.1.1 PAN 2013 Blogs data

The corpus was built using open and public repositories like Netlog with posts having labels about author demographics such as gender and age. The posts were grouped by authors. The age of authors was classified in 3 classes as: 10s (13-17), 20s (23-27), 30s (33-47). The corpus is gender balanced but imbalanced for age groups. The corpus also incorporated a small number of conversations from sexual predators together with samples from conversations between two adults about sex.

Out of total around 225000 samples, for the purpose of this paper's study, I did a random sampling without replacement of 200 samples maintaining the distribution of age and gender population

3.1.2 PAN 2014 Data

This corpus has 4 different sub-corpuses of 4 genres: Social Media, Reviews, Blogs and Twitter Data.

For the purpose of this task I am using only social media and reviews' sub-corpuses.

I didn't include Blogs because it required a heavy amount of parsing skills as the blogs were only half-downloaded giving links for full blog RSS feeds. And those RSS feeds were of different blog providers. So for each provider we would have required different parsers. So this have unnecessarily complicated the task. And also, blogs were included in PAN 2013 data.

I didn't use Twitter either, as tweets were required to be downloaded, but I already had them from PAN 2015 and PAN 2016 data.

3.1.2.1 Social Media Data

The posts were selected from authors having at least 100 average number of words. This subcorpus is balanced by gender but has below distribution for age.

18-24	1550
25-34	2098
35-49	2246
50-64	1838
65+	14

For this subcorpus also, I am performing random sample without replacement of 200 sampled from aroung 7000 data points and preserving the age and gender distribution.

3.1.2.2 Hotel Reviews Data

This subcorpus is derived from another corpus that was crawled from the hotel review site TripAdvisor in the period of one month from mid February to mid March 2009. Short reviews with less than 10 words were removed. The final subcorpus contains around 4000 reviews and has six age classes. Corpus is imbalanced in age but nearly balanced in gender. The distribution shown as below.

Gender	Age	Authors	Reviews
Female	13-17	-	-
	18-24	180	208
	25-34	500	651
	35-49	500	659
	50-64	500	617
	65+	400	494
Male	13-17	-	-
	18-24	180	228
	25-34	500	700
	35-49	500	707
	50-64	500	669
	65+	400	520

I again used only random 200 samples from around 4000 reviews.

3.1.3 PAN-2015 Twitter Data

This corpus contains twitter data with age and gender annotated to the tweets. Age classes were: 1) 18-24, 2) 25-34, 3) 34-49, 4) 50+. The corpus is balanced in terms of gender but skewed because of lower number of users of aged 50 and above.

Distribution is shown below:

18-24	58
25-34	60
35-49	22
50+	12

As this corpus has less than 200 samples, I am using all of them in my experiment.

3.1.4 PAN-2016 Twitter Data

This was built by merging subcorpuses of PAN 2014. So I can use this data as I am not using PAN 2014 twitter data. There are 5 different age classes:

1) 18-24, 2) 25-34, 3) 34-49, 4) 50-64, 5) 65+

The corpus is balanced by gender and age distribution is shown below.

18-24	26
25-34	136
35-49	182
50-54	78
65+	6

I am again using only 200 random samples.

3.2 Preprocessing

- I am randomly subsampling all the data to restrict the sample size to 200. But doing this doesn’t affect the distribution of the data.
- I am extracting raw text from all XML files.
- However, I am not removing any URLs/user mentions or anything such as non-dictionary words, punctuation marks etc.
- I am not processing raw text obtained from parsing because all those pre-processing is inherently handled when I am extracting features using TF-IDF and count vectorizer methods as I am keeping bound on DF.

3.3 Feature Extraction

3.3.1 Stylistic Features

3.3.1.1 POS frequency:

This is the most common style based feature used in almost all studies of Author Profiling. For this feature, I'm extracting POS tags from unprocessed raw text data using NLTK POS tagger. This tagger provides roughly 32 different tags. So I am mapping these into MRC POS tags which has total 10 different tags. The mapping is done as per below rule.

POS Tag	Description	MRC Equivalent
CC	coordinating conjunction	C
CD	cardinal number	O
DT	determiner	O
EX	existential there	*I*
FW	foreign word	O
IN	preposition/subordinating conjunction	R
JJ	adjective	J
JJR	adjective, comparative	J
JJS	adjective, superlative	J
LS	list marker	O
MD	modal	V
NN	noun, singular or mass	N
NNS	noun plural	N
NNP	proper noun, singular	N
NNPS	proper noun, plural	N
PDT	predeterminer	*O*
POS	possessive ending	O
PRP	personal pronoun	U
PRP$	possessive pronoun	U
RB	adverb	A
RBR	adverb, comparative	A
RBS	adverb, superlative	A
RP	particle	R
TO	to	*C*
UH	interjection	I
VB	verb, base form	V
VBD	verb, past tense	V
VBG	verb, gerund/present participle	V
VBN	verb, past participle	V
VBP	verb, sing. present, non-3d	V
VBZ	verb, 3rd person sing. present	V
WDT	wh-determiner	O
WP	wh-pronoun	U
WP$	possessive wh-pronoun	U
WRB	wh-abverb	A

Where, N- Noun, J- adjective, V – Verb, A – Adverb, R-pReposition, C – Conjunction, U – pronoun, I – Interjection, O – Other

I did this for two reasons, to reduce the dimensions of 32 NLTK POS to 10 Common English Language POS, and the other reason was to use MRC Contextual Status of the token.

I build TF vectorizer which takes care or normalization across documents.

3.3.1.2 Word Count, Character Count, Sentence Count

- Word count is inherently handled in POS frequency when I use TF for normalizing.
- I didn't use character count and sentence counts because of the inappropriate format of data. It was not possible to normalize this information, so I discarded them.
- Punctuation mark frequency is also handled in POS.

3.3.1.3 Readability Features

I explored different readability features as they were used in some papers of PAN competition. But I found that of no use because they were not representative of Gender and also for Age, they could have helped for age range of 3 – 18 years. So they can't be used for age prediction task as well.

The papers which used this information didn't quite gain any advantage using these features. Their performance was avg to below avg only.

3.3.1.4 MRC Contexual Status

These features can provide stylistic information about the text. According to this features, a single word can be classified into following categories:

- Specialised,
- Archaic,
- Capital,
- Dialect,
- nonsEnse,
- Foreign/Alien,
- rHetorical,
- erroNeous,
- Obsolete,
- Poetical,
- colloQuial,
- Rare,
- Standard,
- nonce Word

Though I wanted to use these strongly, it turned out to be infeasible because of missing labels in MRC Database and many of the words fell into Standard category. So the distribution turned out to be very skewed and as it would have impacted the machine learning heavily, I didn't use this feature.

3.3.2 Content based features

I wanted to use these features as less as possible because all online social media has little topic bias. Eg. Teens will talk more about sports and gadgets, middle aged people will talk about job, economy, politics. As my project is for cross genre, content based features can't work quite well in formal settings like Formal communication, Newspaper articles, research papers etc.

3.3.2.1 N-gram TFIDF

This was also a common approach by almost everyone in PAN competitions and other studies. I used unigram and bigram as features.

For TFIDF, I am removing 'English' stop words and putting a bound on min and max document frequency to restrict features based on proper token only. I am eliminating tokes having document frequency of >0.5 and <0.5. This gave me appropriate tokens which are actually proper words and found in dictionary. So incorrect words and slangs are eliminated to quite an amount.

3.3.2.2 Other content based features which were explored but not used

- Sentiment Analysis: I did not use this because I thought it would be very biased features for Hotel Review genre. As, most of the reviews will be either positive/neutral. And other genres are more tend to neutral irrespective of age and gender. I should have quantified this assumption but because of lack of proper library for sentiment analysis, I was not able to.
- Topic Modelling using LDA: The same reason why I am avoiding content based features.

3.3.3 IR features

A few participant used IR based features such as Cosine Similarity, and Okapi BM25. This can be implemented using java based library Weka and can't be using sklearn or other Python based open source libraries. However, only before couple of days, I found about python based GraphlabCreate library which implement both of these features and also has support for recommendation systems. But because of time constraint I didn't use this.

4. Experiments

For all the experiments, I am doing training on one type of genre and tasting the model on different type of genre. So I am not performing cross validation or some kind of resampling techniques.

ML algorithms I am using are as follows:

1) Naïve Nayes: It works great with text data and provides really good baseline.
2) Logistic Regression: Works well for binary classification. So using it for Gender prediction only.
3) Random Forest: It provides good accuracy as it's an ensemble technique.

I am evaluating both the content based and stylistic features separately.

In one experiments, I am training on social media data and testing on rest of the genres because social media has highest number of TfIDf features.

In other experiments, I am training on review data and testing on rest of the genres as that data should be topic neutral.

I am also not making any modifications to age category to standardize them across different corpuses because they're quite overlapping and the inconsistency is well handled during finding errors.

5. Results:

Check last pages

6. Conclusion:

It can be seen that both content based and stylistic features works equally on cross-genre gender prediction and performance little more than baseline of 50% accuracy. However, stylistic features work better compared to content based features for age prediction task. Also training genre affects the results in age prediction.

7. Future Work:

This experiment was not able to distinguish between cross genre age and gender prediction. So the future task will to include more such features and also experiment with large sample size and also with other features like IR based features.

8. Acknowledgments

Our sincere thanks to Prof. Markus Dickinson for providing us the required resources to learn and guiding throughout the semester to get a better understanding of the subject.

9. References

1. Madhulika Agrawal and Teresa Gonçalves. Age and gender identification using stacking for classification. In Balog et al.

2. Miguel-Angel 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—notebook for pan at clef 2015.
2015.

3. Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. TEXT, 23:321–346, 2003.

4. Shaina Ashraf, Hafiz Rizwan Iqbal, and Rao Muhammad Adeel Nawab. Cross-genre author profile prediction using stylometry-based approach. In Balog et al. [5].

5. Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors. CLEF 2016 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1609/, 2016.

6. Roy Khristopher Bayot and Teresa Gonçalves. Author profiling using svms and word embedding averages. In Balog et al. [5].

7. Ivan Bilan and Desislava Zhekova. Caps: A cross-genre author profiling system - notebook for pan at clef 2016. In Balog et al. [5].

8. Konstantinos Bougiatiotis and Anastasia Krithara. Author profiling using complementary second order attributes and stylometric features. In Balog et al.
[5].

9. John D. Burger, John Henderson, George Kim, and Guido Zarrella.
Discriminating gender on twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
7 http://scikit-learn.org//
8 http://www.adobe.com/

10. Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors. CLEF
2014 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings
(CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/ Vol-1180/, 2014.

11. Rodwan Bakkar Deyab, José Duarte, and Teresa Gonçalves. Author profiling
using support vector machines. In Balog et al. [5].

12. Daniel Dichiu and Irina Rancea. Using machine learning algorithms for author profiling in social media. In Balog et al. [5].

13. Pamela Forner, Roberto Navigli, and Dan Tufis, editors. CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain, 2013.

14. Pepa Gencheva, Martin Boyanov, Elena Deneva, Preslav Nakov, Georgi Georgiev,Yasen Kiprov, and Ivan Koychev. Pancakes team: a composite system ofdomain-agnostic features for author profiling. In Balog et al. [5].

15. Tim Gollub, Benno Stein, and Steven Burrows. Ousting ivory tower research:
towards a web framework for providing experiments as a service. In Bill Hersh,Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, 35th International
ACM Conference on Research and Development in Information Retrieval (SIGIR 12), pages 1125–1126. ACM, August 2012. ISBN 978-1-4503-1472-5.

16. 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. ISBN
978-1-4673-2621-6.

17. Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Recent trends in digital text forensics and its evaluation. In Pamela Forner, Henning Müller, Roberto Paredes,

Paolo Rosso, and Benno Stein, editors, Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative (CLEF 13), pages 282–302, Berlin Heidelberg New York, September 2013. Springer. ISBN 978-3-642-40801-4.

18. Janet Holmes and Miriam Meyerhoff. The handbook of language and gender. Blackwell Handbooks in Linguistics. Wiley, 2003.

19. Mirco Kocher and Jacques Savoy. Unine at clef 2016: author profiling. In Balog et al. [5].

20. Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. literary and linguistic computing 17(4), 2002.

21. H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In 25th International Conference on Machine Learning pp. 536–543. ICML'08, ACM, 2008.

22. A. Pastor Lopez-Monroy, Manuel Montes-Y-Gomez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Esau Villatoro-Tello. INAOE's participation at PAN'13: author profiling task—Notebook for PAN at CLEF 2013. In Forner et al. [13].

23. A. Pastor López-Monroy, Manuel Montes y Gómez, Hugo Jair-Escalante, and Luis Villase nor Pineda. Using intra-profile information for author profiling—Notebook for PAN at CLEF 2014. In Cappellato et al. [10].

24. Suraj Maharjan, Prasha Shrestha, Thamar Solorio, and Ragib Hasan. A straightforward author profiling approach in mapreduce. In Advances in Artificial Intelligence. Iberamia, pages 95–107, 2014.

25. Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, and Alexander Gelbukh. Adapting cross-genre author profiling to language and corpus. In Balog et al. [5].

26. Pashutan Modaresi, Matthias Liebeck, and Stefan Conrad. Exploring the effects of cross-genre machine learning for author profiling in pan 2016. In Balog et al. [5].

27. Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "how old do you think i am?"; a study of language and age in twitter. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013.

28. Eric W. Noreen. Computer intensive methods for testing hypotheses: an introduction. Wiley, New York, 1989.

29. James W. Pennebaker. The secret life of pronouns: what our words say about us. Bloomsbury USA, 2013.

30. James W. Pennebaker, Mathias R. Mehl, and Kate G. Niederhoffer. Psychological aspects of natural language use: our words, our selves. Annual review of psychology, 54(1):547–577, 2003.

31. Oliver Pimas, Andi Rexha, Mark Kroll, and Roman Kern. Profiling microblog authors using concreteness and sentiment - know-center at pan 2016 author profiling. In Balog et al. [5].

32. Francisco Rangel and Paolo Rosso. On the multilingual and genre robustness of emographs for author profiling in social media. In 6th international conference of CLEF on experimental IR meets multilinguality, multimodality, and interaction, pages 274–280. Springer-Verlag, LNCS(9283), 2015.

33. Francisco Rangel and Paolo Rosso. On the impact of emotions on author profiling. Information processing & management, 52(1):73–92, 2016.

34. Francisco Rangel, Paolo Rosso, Moshe Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. In Forner P., Navigli R., Tufis D. (Eds.), CLEF 2013 labs and workshops, notebook papers. CEUR-WS.org, vol. 1179, 2013.

35. 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 Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 labs and workshops, notebook papers. CEUR-WS.org, vol. 1180, 2014.

36. Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd author profiling task at pan 2015. In Cappellato L., Ferro N., Jones G., San Juan E. (Eds.) CLEF 2015 labs and workshops, notebook papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391, 2015.

37. 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.

38. 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, pages 199–205. AAAI, 2006.

39. Ma. José Garciarena Ucelay, Ma. Paula Villegas, Dario G. Funez, Leticia C. Cagnina, Marcelo L. Errecalde, Gabriela Ramírez-De-La-Rosa, and Esau Villatoro-Tello. Profile-based approach for age and gender identification. notebook for pan at clef 2016. In Balog et al. [5].

40. Ben Verhoeven and Walter Daelemans. Clips stylometry investigation (csi) corpus: a dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In 9th International Conference on Language Resources and Evaluation (LREC 2014), 2014.

41. Ben Verhoeven, Walter Daelemans, and B. Plank. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In 10th International Conference on Language Resources and Evaluation (LREC 2016), 2016.

42. Mart Busger Op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva,

Chris Pool, Johannes Bjerva, Hessel Haagsma, and Malvina Nissim. Gronup: Groningen user profiling. In Balog et al. [5].

43. Edson Weren, Anderson Kauer, Lucas Mizusaki, Viviane Moreira, Palazzo de Oliveira, and Leandro Wives. Examining multiple features for author profiling. In Journal of Information and Data Management, pages 266–279, 2014.

44. Alexander Yeh. More accurate tests for the statistical significance of result differences. In Proceedings of the 18th Conference on Computational Linguistics - Volume 2, pages 947–953, Stroudsburg, PA, USA, 2000. Association for
Computational Linguistics.

45. Cathy Zhang and Pengyu Zhang. Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA, 2010.