Final Report on

An approach for Author Profiling on data from heterogeneous social media sources

CS 585, UMass Amherst, Fall 2017

Submitted By:

Sanjay Reddy S (ssatti@umass.edu)

Aditya Agrawal (adityaagrawa@cs.umass.edu)

College of Information and Computer Sciences

University of Massachusetts, Amherst

MA

Table of Contents

Abstract

Author profiling includes the study of various profiling traits of an author such as age and gender. This work describes our methodology for the task of cross-genre author profiling, inspired from the PAN 2016 challenge. Gender and age prediction problem is addressed as a classification task and approach to this problem focuses on feature engineering. Furthermore, in our final model we combine different models which work well with different types of social media data using minimal, simplistic features. With our approach, we are able to achieve competitive accuracies for gender classification task and an improvement in the age classification task, as we further modified the problem to a more viable and useful task.

1. Introduction

In this study, we try to gain an understanding of how authors of different classes (e.g., old men and young women) possess distinct characteristics while authoring text and which textual features might characterize people in the same class. For example, younger people may be more prone to making spelling errors than older people. Author profiling distinguishes between classes of authors studying their *sociolect* aspect, that is, how language is shared by people. This helps in identifying profiling aspects such as gender, age, native language, or personality type based on the topics and the linguistic features present in a person's writings and authored content. Author profiling has a lot of applications in the real world and is a prominent task in natural language processing.

This study would find application in the fields such as security, forensics and marketing, as studying in detail various profile characteristics of an author has gained significance in the contemporary times. Especially, the PAN [1] (Uncovering Plagiarism, Authorship, and Social Software Misuse) competition garnered attention in the task of author profiling as a part of the CLEF conference since 2013.

Author profiling becomes useful when there is missing information about authors which could potentially be relevant for an organization. For instance, advertising campaigns could be enhanced if an organization has demographical information about its target customer base- information like likes/dislikes from their online product reviews. Domains such as abstractive text summarization could make use of author profiling techniques by differentiating between human-written and machine-generated summaries. Yet another application of social media author profiling could be in cyber-crimes investigation- one would like being able to know the linguistic profile of the author of a harassing text message (language used by a certain type of people) and identify certain characteristics (language as evidence).

Two profile aspects- namely *age* and *gender*, have been the cynosure for the PAN author profiling competitions. PAN 2016 challenge focused on the shared task [3] of cross-genre age and gender

identification. This entailed training the documents on one genre (Twitter, blogpost etc.) and the evaluation on another (unknown to the participants prior to software submission) genre, such as blogs or other forms of social media data (ranging from one-liner tweets, to paragraphs of blogposts.). English, Spanish and Dutch were the languages that were provided in the 2016 challenge. The Dutch language data was very limited, as was called out by many participants which lead to unfairly poor results on the same [9].

In this project, we present different approaches for various sub-tasks in author profiling and compare efficiencies of various algorithms by running them on a dataset. The focus is on author profiling in social media since we are mainly interested in everyday language and how it reflects basic social and personality processes. We plan on tackling the cross-genre author profiling task as a classification problem using different models like- k-Nearest Neighbor, Decision Trees, SVM, Logistic Regression (and/or ensembles), mainly focusing our research on feature engineering- for creating a language-independent cross-genre (across different forms of formal and informal writing) author profiling model of high accuracy. We plan on combining the efforts of the top winners of PAN 2016 challenge and to use the research done over PAN 2014-16 to improve upon the accuracy achieved over this time period. PAN challenges have lately focused on author profiling tasks- shifting from domain specific profiling tasks to cross-genre author profiling, and the major reason for fairly low accuracies, even amongst the best of submissions, is attributed to the difficulty of the cross-genre nature of author profiling where the source platform of the data is unknown and the model is expected to generalize over the subtle differences between these platforms. Our approach herein would be to create a feature-focused simplistic model of high accuracy which generalizes well to any social media platform data, which concentrates on the major features which occur across different languages, instead of using high number of language-specific features- which would involve a trade-off of accuracy and generic nature of our model.

2. Problem Statement

To create a cross-genre author profiling model which for an input of any form of social media data- ranging from one-line tweets to paragraphs of a blog, outputs the gender and age category of the text, which uses minimal features, simple design, and furnished high accuracy output.

3. Related Work

3.1 S. Argamon *et al* [2]

This work shows how the right combination of linguistic features and machine learning methods enables an automated system to effectively determine several such aspects of an anonymous author. This work serves as a very good introduction to the whole problem of author profiling. Basic outcomes of such a system like age, gender, native language, and personality are explained in great detail, and a general roadmap of how such systems are usually built is also presented.

3.2 Overview of PAN 2016 [3]

This work presented the framework and the results of the Author Profiling task at PAN 2016. The objective was to predict age and gender from a cross-genre perspective. For this purpose a corpus from Twitter was provided for training, and different corpora from social media, blogs, essays, and reviews was provided for evaluation. Altogether, the approaches of 22 participants were evaluated. They also explain the training datasets given, testing data used and also the evaluation metrics upon which the teams are being ranked. It also provides us reference to the submission of individual teams.

3.3 Overview of PAN 2014 [4]

This work presented the framework and the results for the Author Profiling task at PAN 2014. Task of 2014 was to analyze the adapt-ability of the detection approaches when given different genres. For this purpose, a corpus with four different parts (sub corpora) was compiled: social media, Twitter, blogs, and hotel reviews. Altogether, the approaches of 10 participants were evaluated. It also explains the training datasets given, testing data used and also the evaluation metrics upon which the teams were ranked. This work provided us with some of the pre-annotated data for the different domains we wanted to work on.

3.4 Overview of PAN 2013 [5]

This work presented the framework and results for the Author Profiling task at PAN 2013. It described in detail the corpus and its characteristics, and the evaluation framework that was used to measure the participants performance to solve the problem of identifying age and gender from anonymous texts. Finally, the approaches of the 21 participants and their results are described. This work was another source for our pre-annotated data.

3.5 Busger et al - PAN 2016 [6]

Busger et al trained a SVM linear model on tweets to perform user profiling, in terms of gender and age, on non-Twitter social media data, whose actual nature was unknown at development. Results on test data were lower than what was normally observed for this task — that is, when done in a single-genre setting — but were in fact the state-of-the-art for the cross-genre setting at PAN 2016. It was the top work at PAN 2016 and provisioned us a basic overview of author profiling and about what features were used to train their SVM model. It also provided us extensive experimentation results, by modifying various features they used in training. This work provided us a decent benchmark for improvisation and served as a source for motivation.

3.6 Miguel-Angel et al - PAN 2015 [7]

This work uses dimensionality reduction techniques on the top of typical discriminative and descriptive textual features for author profiling task. The main idea is that each representation, using

the full feature space, automatically highlights the different stylistic and thematic properties in the documents. Specifically, it proposes the joint use of Second Order Attributes (SOA) and Latent Semantic Analysis (LSA) techniques. This work discussed other features and models which can be used when author profiling is specifically applied to social media data and thus, provided an instrument for enhancing our domain knowledge.

3.7 F. Rangel et al – PAN@FIRE [8]

This work at PAN@FIRE track on Personality Recognition from SOurce COde (PR-SOCO) has addressed the problem of predicting author's personality traits from her source code. In this paper, F. Rangel et al analysed 48 runs sent by 11 participant teams. Given a set of source codes written in Java by students who answered also a personality test, participants had to predict personality traits, based on the big five model. This work brings forth an interesting extension of author profiling, where in, personality traits are extracted from just using the source code written by the user, which does not directly relate to our domain target, but allows for extension of the horizon of our approaches.

3.8 Goswami et al – Stylometric Analysis of Bloggers' Age and Gender [9]

This work reported results of stylometric distinctions in blogging for gender and age group variation. It was based on two mutually independent features. Two main features which were in focus were- the use of slang words and also the average word length per sentence for different age groups. The combined feature list improved the accuracy by a significant extent in predicting age and gender. These experiments were conducted on a 20,000 blog corpus. This work gave us insights in handling with blogpost data and for integrating age distinction for really large texts in our author profiling model. We have also used average words per sentence for different author categories for aiding our results.

3.9 P. Modaresi et al – PAN@FIRE [10]

This work described the approach proposed by Modaresi et al for the PAN 2016 Author Profiling challenge. They extracted stylistic and lexical features for training a logistic regression model, which became a mainstay for our work as well. It further, gave us insights into why the cross-genre nature of the task makes the task really difficult to solve, by discussing the effects in detail. This work was the winner for the gender detection task and joint winners for the other tasks.

3.10 A. Pastor et al – PAN 2013 [11]

This work described the participation of the Laboratory of Language Technologies of INAOE at PAN 2013 evaluation lab. They adopted second-order attribute representations to overcome standard issues of the Bag of words representations. They utilized document vectors for achieving the same. This work motivated us for using second-order attributes and thus, we have included the same in the proposed work section for future scope.

4. Data

We have proposed to use the datasets provided by PAN [1] and we are moving forward with it. We have, further, only focused on the English language datasets. In specific, we primarily started out using year 2014 because the dataset was conveniently sub-divided into various categories like blogs, reviews, social-media and twitter.

A) English blogs: 147 files
B) English Reviews: 4,167 files
C) English social-media: 7,746 files

Further, we have sourced our data from the following PAN challenges for different types of data:

1. PAN 2013 [5]: Conversations dataset
2. PAN 2014 [4]: Hotel Reviews & Twitter datasets
3. PAN 2016[3]: We used this for reference, but did not directly include the twitter data as there was major overlap in 2014 and 2016 twitter data [9].
4. PR-SOCO[8]: We used this data for initial exploratory analysis into the domain of author profiling. (Work has been summarized in the Methodology section.)

All these categories also have a ground truth .txt file, which specifies the gender and also the age-bucket into which these authors fall into. Further, it was not possible to obtain the test data for the PAN 2016 challenge as the data was password protected. Additional statistics about the characters in data have been mentioned below, from our initial exploratory data analyses, where the metrics report the number of characters per item in each dataset:

Characters description for Twitter Data (Tweets):

Max Length	Min Length	Median Length	1st Quartile	3rd Quartile	Average Length:
140	1	66.0	41.0	95	67.6711808724

Character Descriptions for Conversation Dataset:

Max Length	Min Length	Median Length	1st Quartile	3rd Quartile	Average Length:
63927	3	102.0	53.0	221.0	293.730751063

Characters description for Hotel Reviews:

Max Length	Min Length	Median Length	1st Quartile	3rd Quartile	Average Length:
16673	8	788.0	387.0	1270.25	1009.87160675

All the files were .xml files and we wrote down a basic framework in python which parsed the xml document objects and creates the necessary .txt files which we were used in our analyses. Also, we were successful in getting hold of Personality Recognition in SOurce code (PR-SOCO)[7] data and experimented with the same initially, to learn more about the nuances of the task of author profiling. Further, some of the characteristics and statistics of our datasets under consideration have been mentioned below:

- Gender Distribution for Conversations: {'MALE': 21346, 'FEMALE': 10409}
- Gender Distribution for Twitter datasets: {'MALE': 14192, 'FEMALE': 13203}
- Gender Distribution for Hotel Reviews: {'MALE': 2823, 'FEMALE': 2629}
- Age Distribution for Conversations: {'18-24': 1072, '23-50': 30683}
- Age Distribution for Tweets: {'18-24': 1878, '23-50': 19083, '50-xx': 6434}
- Age Distribution for Hotel Reviews: {'18-24': 436, '23-50': 2717, '50-xx': 2299}
- Age Distribution for Conversations (Reduced buckets): {'13-17': 1072, '33-37': 12261, '23-27': 18422}
- Age Distribution for Tweets (Reduced buckets): {'18-24': 1878, '35-49': 11292, '50-64': 5807, '25-34': 7791, '65-xx': 627}
- Age Distribution for Hotel Reviews (Reduced buckets): {'35-49': 1366, '18-24': 436, '25-34': 1351, '50-64': 1285, '65-xx': 1014}

5. Methodology

All the leading teams in PAN competitions (across the years) have used SVM, Logistic Regression and various other machine learning models to achieve good results. We wanted to use the features developed and used across the years (that is, to learn from past winners as a whole) and then use combinations of various approaches to merge and provision an improved form of author profiling.

One experimentation, which we performed during the exploratory analysis phase involved trying personality recognition from source code of a programmer, for the sake of understanding more about author profiling and its subtleties as a task. In PR-SOCO we were given code written by users and we had to predict the personality of the coder. In the dataset provided, Personality was defined using five traits (from the Big Five Theory): extroversion (E), emotional stability / neuroticism (S), agreeableness (A), conscientiousness (C), and openness to experience (O). By going through the dataset, we figured that the task was too tough (at least at this point in time).

We approached this problem in a similar style as with other datasets and first started on Feature engineering. But in code datasets, features like POS tags are out of question and also, traditional vocabulary statistics won't be helpful/present. For example, in gender we can

probably analyze the types of words used and make a prediction that men use more action words and women use more expressive & flowery language. In fact, in normal mood and personality detection systems, the words used, play an important role. But in programs, the basic vocabulary and syntax is by definition fixed and very constrained. The keywords remain the same irrespective of the coder. Thus, getting features was exceedingly complicated. Coupled with the fact that the dataset had very few files (code written by just 50 programmers) made this task all the more tough. That being said, we did decide on few features like: number of variables, number of functions, number of comments, length of variables used, length of comments typed etc. But even that isn't too perfectly done and ultimately, we scraped off the idea completely. We pushed this out to future scope, as this is one area which we can work on (primarily because if it gives good accuracy it might be very helpful for companies).

For our main model, we started off by performing the extraction of our datasets (which were in .xml formats) from the various sources discussed above using our framework and then proceeded to the pre-processing of the dataset, removing and handling the elements mentioned below. Most of these pre-processing steps has been inspired by the word done by previous submissions at the PAN challenges, where we have tried to combine the best set of pre-processing steps:

Preprocessing-

- Modified urls to 'URL' and numbers to 'NUM'.
- Removed html tags and @'at-the-rate'mentions.
- Eliminated duplicate 'retweet' data, which would have added undue bias to our dataset.
- Eliminated noisy tweets containing non-Latin characters.
- Removed accents from our dataset.

Next, we proceeded to the creation of feature vectors for our data and concentrated on the following general features, which we gleaned out of our proposed:

Features vector elements for our dataset:

1. Total characters per item.
2. Total words per item.
3. Counts of different Part-Of-Speech tags per item.
4. Starts with capital (To capture difference between: The man is good. & the man is good.)
5. Percentage of capitalized tokens per sentence (Example: The United States of America.)
6. Capital words per sentence. (Example: THE UNITED STATES OF/ LOLLLL)
7. Ending of sentence with a punctuation. (To capture difference between- it was bad & It was bad.)
8. Percentage of punctuations percentage per sentence. (To capture- lol !!!!!!!!!!)
9. Average word length per sentence in our given item.
10. Percentage of Out-of-dictionary words (Using enchant library [13]) or even slang words.
11. F-Measure for Part-Of-Speech Tags.
12. Count of the number of Hashtags.

Libraries like nltk and pyenchant were used for obtaining the features. Further, we wanted to develop a model which depending upon the length of input decides the internal model to use (from Decision Tree, Logistic Regression, or SVM, etc.) and will according to that selectively activate different set of features. This lead to us dividing our any form of social media data into the following 3 broad categories:

1. Tweets/ One-liners (140 characters or less)
2. Conversation Data (Intermediate – could be a facebook post or Instagram caption)
3. Review Data (This is the longest data available in our category which could be a blog or review posts.)

We considered that a longer length may show a blog is being analyzed and some features would weigh in more for such classifications.
Finally we wanted to look at the different model(s) which would work for a given type of social media data and this lead us to experimenting with the following machine learning models, which included extensive feature engineering, on models which have been used over the years by submissions at the PAN challenges. The models we considered for our experimentation are given below:

1. K-Nearest Neighbor
2. Gaussian Naïve Bayes
3. Logistic Regression
4. Support Vector Machines
5. Decision Trees
6. Voting Classifiers (Ensemble- KNN, Logistic Regression, and SVM)
7. Voting Classifiers (Ensemble- KNN, Decision Tree, Logistic Regression, and SVM)

The above models were implemented in sklearn and graphs plotted using matplotlib. Additionally, a 90-10 split was utilized for training and testing the data. The feature engineering that was performed for learning the best subset of features from our superset for different types of social media data has been discussed in detail in the next section.

Finally, after deciding upon our set of machine learning models which individually performed better for different categories of data and the subset of features per data category, we went onto implementing the complete unified model which integrated our learning, offering us the results provided in the Results section.

5.1 Pre-existing Software Systems which were used:

(a) Language used: Python2.7
(b) Word relations library: nltk, pyenchant.
(c) Machine Learning utilities in Python: sklearn, numpy, matplotlib etc.

6. Feature Engineering

After studying the previous submission at the PAN 2016 challenge we realized that feature engineering would be of utmost important for this task and a powerful model would be possible if we are able to efficiently generalize our model using the minimum number of features as possible.

(Modaresi *et al* [9] worked with only 5 features on logistic regression and got similar results as Busger *et al* [5], where 10 odd features were used with a linear Support Vector Machine model.) We started off by considering around 30 features initially, but as we dived deeper into implementations of some of the features, the complicated nature of those features came to the fore and finally we zeroed down to the set of features mentioned in the previous section.

Next, we wanted to experiment with various possible combination of these features, i.e. subsets of our feature set with different learning models, for every type of data we had. We experimented with the following subsets of our features, the results of which are presented in the results section of this report:

Different Groups combinations of features experimented upon (per item):

Grp1. Total words, total characters, count of hashtags, average word length, percentage of punctuations, percentage of wrong words, total capital words, item has capitalized text, count of hashtags.

Grp2. Total words, count of hashtags, average word length, percentage of punctuations, percentage of wrong words, count of hashtags.

Grp3. Total words, count of hashtags, average word length, percentage of punctuations, percentage of wrong words, count of hashtags, total capital words.

Grp4. Total words, count of hashtags, average word length, percentage of punctuations, percentage of wrong words, count of hashtags, part of speech statistics.

Grp5. Total words, count of hashtags, average word length, percentage of punctuations, percentage of wrong words, count of hashtags, part of speech statistics, total capital words.

Grp6. Total words, count of hashtags, average word length, percentage of punctuations, percentage of wrong words, count of hashtags, part of speech statistics, total capital words, item has capitalized text.

Grp7. Number of words, number of characters, f-measure for part of speech statistics, counts of incorrectly spelled words, item starts with a capital letter, item has capitalized text, count of hashtags, percentage of incorrectly spelled words, percentage of punctuations, average word length, item ends with a period, total capital words, part of speech statistics.

Grp8. Number of words, number of characters, f-measure for part of speech statistics, counts of incorrectly spelled words, item starts with a capital letter, item has capitalized text, count of hashtags, percentage of incorrectly spelled words, percentage of punctuations, average word length, item ends with a period, total capital words, part of speech statistics, number of punctuations.

Grp9. Number of words, number of characters, f-measure for part of speech statistics, counts of incorrectly spelled words, item starts with a capital letter, item has capitalized text, count of hashtags, percentage of incorrectly spelled words, percentage of punctuations, average word length, item ends with a period, total capital words.

Grp10. Number of words, number of characters, f-measure for part of speech statistics, item starts with a capital letter, item has capitalized text, count of hashtags, percentage of incorrectly spelled words, percentage of punctuations, average word length, item ends with a period, total capital words, part of speech statistics.

Grp11. Number of words, number of characters, f-measure for part of speech statistics, item starts with a capital letter, item has capitalized text, count of hashtags, percentage of incorrectly spelled words, percentage of punctuations, average word length, item ends with a period, total capital words, part of speech statistics, number of punctuations.

Grp12. Number of words, number of characters, f-measure for part of speech statistics, item has capitalized text, count of hashtags, percentage of incorrectly spelled words, percentage of punctuations, average word length, item ends with a period, total capital words, part of speech statistics, number of punctuations.

Grp13. Number of words, number of characters, f-measure for part of speech statistics, item starts with a capital letter, item has capitalized text, count of hashtags, percentage of incorrectly spelled words, percentage of punctuations, average word length, total capital words, part of speech statistics, number of punctuations.

Finally, accuracies for these different feature subsets using different machine learning models were computed and plotted using sklearn and matplotlib respectively. These have been presented in our results section.

7. Results

The results have been discussed in this section, starting off with the feature engineering. We selected the above mentioned subset groups of features for our model and wanted to find the best set of features which in combination with different models, which would yield our best accuracy. In general we found that the following models worked well for the following datasets:

1. Twitter dataset- Decision Tree classified performed better than other models.
2. Conversation dataset – Logistic regression performed better than other models.
3. Reviews dataset – Logistic regression performed better than other models.

Further, when we plotted graphs for our different groups of subsets out of our superset of features, we obtained the following, where the group numbers correspond to the group numbers mentioned in the previous section:

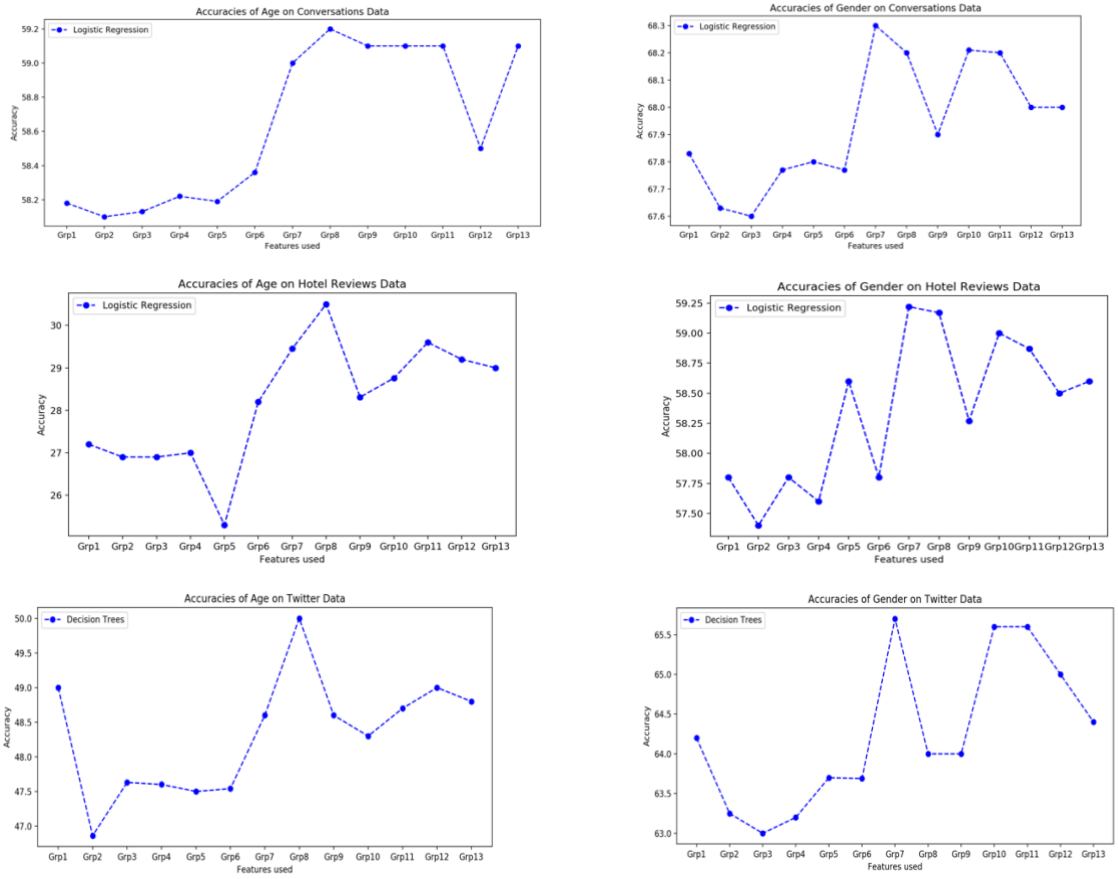

Fig 1. Accuracy Comparisons for different types of data against different groups of features. (Here the X-axis corresponds to the GRPX as discussed in previous Feature Engineering Section.)

Finally, after we had decided on our different feature subsets, per data type and the different model to use, we ran our model and tested the same. Further, we wanted to narrow down our prediction for the age category as earlier the age category was divided into the following 6 bins:

['18-24' '25-34' '35-49' '50-64' '65-xx']

We decided that such fine distinction between age is not very feasible and a narrower focus would give finer and accurate results, which would be more useful. Knowing whether the object is a child, adult, or old with a higher accuracy is more useful, than looking for very close approximate with a bad accuracy. So, we transposed our data into the following 3 bins for the age group:

['18-24' '23-50' '50-xx']

Before we present our accuracies, we'll include below the accuracies from respective challenges for the different types of datasets under our consideration:

English			
Team	Total	Gender	Age
Ladra	0.3301	0.5631	0.5924
Gillam	0.3245	0.5413	0.5947
Jankowska	0.2796	0.5185	0.5463
baseline	0.1649	0.4997	0.3324
Aleman	0.0162	0.0277	0.0278

Fig 2. Accuracies for Conversation Data

English			
Team	Joint	Gender	Age
lopezmonroy14	0.5333	0.7667	0.6333
shrestha14	0.4000	0.7333	0.4333
liau14	0.3667	0.6667	0.5667
marquardt14	0.3000	0.5667	0.5333
baker14	0.2667	0.5333	0.5000
ashok14	0.2333	0.5000	0.4667
castillojuarez14	-	-	-

Fig 3. Accuracies for Twitter Data

English			
Team	Joint	Gender	Age
liau14	0.2622	0.7317	0.3415
lopezmonroy14	0.2500	0.6524	0.3720
shrestha14	0.2012	0.6280	0.2805
marquardt14	0.1585	0.5976	0.2561
ashok14	0.1220	0.5854	0.2317
baker14	0.1037	0.5427	0.2439
castillojuarez14	0.0854	0.4756	0.1951

Fig 3. Accuracies for Review Data

Contrasting with above given PAN accuracies, one shall view our accuracies for our different datasets and different categories below, including both 6 bins and 3 bins distinction for the age group:

Accuracy	Review Data	Twitter Data	Conversation Data
For Gender	59.09 %	66.43 %	67.79 %
For Age (For 6 bins)	27.20 %	48.5 %	58.36 %
For Age (For 3 bins)	54.7 %	71.2 %	96.6%

Finally, before we present our final joint accuracies (joint of all models) for age and gender, we include below sample accuracies from the PAN Challenges, which help contrast our accuracies which are fair for the gender classification task, but performs really well for the age classification task.

English			
Team	Joint	Gender	Age
Waser*	**0.2098**	0.5230	**0.3879**
Busger *et al.*	0.1897	**0.5575**	0.3046
Devalkeneer	0.1839	0.5259	0.2931
Dichiu & Rancea	0.1753	0.5345	0.2989
Agrawal & Gonçalves	0.1724	0.5431	0.3103
Bougiatiotis & Krithara	0.1724	0.5345	0.3046
Modaresi(a)	0.1724	0.5057	0.3218
Bilan *et al.*	0.1667	0.5374	0.2902
Gencheva *et al.*	0.1638	0.5287	0.2902
Garciarena *et al.*	0.1609	0.5201	0.2816
Kocher & Savoy	0.1552	0.5144	0.2816
Modaresi *et al.*	0.1552	0.5029	0.3017
Zahid	0.1523	0.4885	0.3103
Ashraf *et al.*	0.1494	0.4971	0.2902
Roman-Gomez	0.1494	0.5144	0.2874
Bakkar *et al.*	0.1466	0.5029	0.2874
baseline	0.1207	0.5402	0.2126
Pimas *et al.*	0.0057	0.0201	0.0086

Fig 4. Joint accuracy at PAN 2016

Accuracy	With 6 bins of Ages	With 3 bins of Ages
For Gender	53.775%	52.9557%
For Age	32.930%	54.735%

Table. Our accuracies for heterogeneous data.

8. Future Scope

The datasets primarily deal with age and gender, though from differing sources like tweets and documents. The current scope of the project was two-fold: obtaining good accuracy by improvising over different models and also to observe how well the same models work when the domain is

changed (for example, discerning how well models trained on twitter data work on something disparate like restaurant reviews). We further would like to explore the use of second-order attributes in our unified classification model, expand the horizon of the task by observing the performance of our model on different language datasets and maybe incorporate sentiment analysis for using a sentiment indicator as another feature in our model. Going forward we would like to take this work and make it language independent, as we have avoided complex features as far as possible and have been able to reproduce reasonable accuracies and even better for the age classification task.
Additionally, we can use this approach for other classifications. For Example, PAN 2013 is centered around conversations which include few from sexual predators also. These sorts of valuable insights can also be gained using these models. Although we have extracted the data files, we haven't written files specifically for criminal classification because we didn't have any baseline to compare against. But the models are written generically enough that adding and testing would be easy enough.
Further, modern and state-of-the-art models like deep neural networks can also be explored, which have been purveying great results, but with an increased computation overhead. The usual consensus though for the use of the same has been when there is an abundance of features.

9. Conclusion

We have presented our approach and model for the cross-genre author-profiling task for social media data in English language, which was inspired from the PAN 2016 challenge. Our best results for the gender and age classification tasks in terms of accuracy are 53.775% and 32.930% across 6 age categories for English data and 52.9557% for gender and 54.735% for 3 bin age category. The age classification model is of fairly high accuracy, which competes amongst the top 10 accuracy spots across all PAN 2016 challenges, although the gender model performs just slightly better than using chance. We attribute this poor result for the gender classification task to the inherent complicated nature of the gender classification problem, where obtaining distinctions or rather, existence of distinctions between writings by men and women is still a hotly debated topic in the world [x- cite article about impossibility of this problem]. Additionally, the training set for age was imbalanced, which resulted in comparatively poor performance than what we expected. Techniques such as sampling or SMOTE could have been used to overcome this problem[9].

During the experimentation phase, we tried different feature combinations, with different machine learning models and explored the same. Searching for good genre-independent features was a difficult task, so we settled for a trade-off in accuracy and generalization. In our future work, we will include more language-independent features to better capture the characteristics of each language and tackle the second part of the PAN 2016 challenge, i.e.- creating a language-independent cross-genre author profiling model.

10. References

[1] "PAN", Pan.webis.de, 2017. [Online]. Available: http://pan.webis.de/index.html. [Accessed: 23- Dec- 2017].

[2] S. Argamon, M. Koppel, J. Pennebaker and J. Schler, "Automatically profiling the author of an anonymous text", *Communications of the ACM*, vol. 52, no. 2, p. 119, 2009.

[3] Balog K., Capellato L., Ferro N., Macdonald C. (Eds.) "Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations", CLEF 2016 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1609, pp. 750-784.

[4] Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) "Overview of the 2nd Author Profiling Task at PAN 2014", CLEF 2014 Labs and Workshops, Notebook Papers. CEUR-WS.org, vol. 1180, pp. 898-827.

[5] Forner P., Navigli R., Tufis D. (Eds.) "Overview of the Author Profiling Task at PAN 2013", Notebook Papers of CLEF 2013 LABs and Workshops. CEUR-WS.org, vol. 1179.

[6] 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.

[7] 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.

[8] F. Rangel, F. González, F. Calle, M. Montes and P. Rosso, "PAN at FIRE: Overview of the PR-SOCO Track on Personality Recognition in SOurce COde", 2017.

[9] Goswami S., Sarkar S., Rustagi M., Stylometric Analysis of Bloggers' Age and Gender. International AAAI Conference on Web and Social Media, North America, mar. 2009.

[10] Modaresi, Pashutan et al. "Exploring the Effects of Cross-Genre Machine Learning for Author Profiling in PAN 2016." CLEF (2016).

[11] "Do women and men write differently?", language: a feminist guide, 2017. [Online]. Available: https://debuk.wordpress.com/2016/03/06/do-women-and-men-write-differently/. [Accessed: 19- Nov- 2017].

[12] A. Pastor, M. Gomez, H. J.Escalante, L.V.Pineda, and E.V.Tello. INAOE's Participation at PAN'13: Author Profiling task—Notebook for PAN at CLEF 2013. In Forner et al.

[13] PyEnchant,2017.[Online].Available:http://pythonhosted.org/pyenchant/api/enchant.html.

Link to Final Code- https://1drv.ms/u/s!AtchQeadaB7FgVCHf0n33fbQV-hs