Author Profiling from Personal Content Blogs

Aayush Singhal
12CS10002

Aseem Patni
12CS10008

Soham Dan
12CS10059

Bhushan Kulkarni
12CS30016

Pranay Yadav
12CS30025

Shubham Saxena
12CS30032

Sruthi Warrier
Mentor

Anurag Verma
Mentor

Suman Kalyan Maity
Mentor

ABSTRACT

This project aims to predicting personally identifiable information (PII), such as age and gender of the author by extracting features from his/her personal content blog texts. We intend to define the state-of-the-art in the field and overcome the shortcomings of the prior works in the personality recognition tasks. This report is meant to share our progress so far and contains details about our future plans.

Keywords

Authorship Profiling, PII, Blogosphere, Natural Language Processing

1. INTRODUCTION

Though the enormous impact of social media on our daily life, we observe a lack of information about those who create the contents. In this regard, author profiling tries to determine the gender, age, native language or personality type of authors by analyzing their published texts. In this study, we focus on building a system to identify only the gender and age of the authors. Other authorship details will be a part of the future work in this area. Author profiling is of growing importance: E.g., from a marketing viewpoint, companies may be interested in knowing the demographics of their target group in order to achieve a better market segmentation; from a forensic viewpoint, determining the linguistic profile of a person who wrote a ”suspicious text” may provide valuable background information.

This study is targeted towards partial fulfillment of requirements for *CS60057: Speech & Natural Language Processing* during Fall 2015, under the guidance of Prof. Pawan Goyal.

The remainder of this paper is organized as follows. Section 2 describes the corpus, Section 3 covers the proposed approach, Section 4 presents the results obtained so far, Section 5 discusses the evaluation measures, Section 6 contains details about the future work-plan and Section 7 enumerates the references. Please find attached PDF containing details about the work done by the individual team-mates.

2. DATA-SET

2.1 Corpus

We have used the following two corpora for this study:

- **Blog Authorship Corpus [6]**
  The Blog Authorship Corpus consists of the collected posts of 19,320 bloggers gathered from blogger.com in August 2004. The corpus incorporates a total of 681,288 posts and over 140 million words or approximately 35 posts and 7250 words per person. All bloggers included in the corpus fall into one of three age groups — ”10s” [13-17], ”20s” [23-27], ”30s” [33-47]. For each age group there are an equal number of male and female bloggers.
- **PAN’14 Corpus**
  As a part of the Author Profiling Shared Task in PAN ’14, this corpus was made available for use during the competition. This data-set originally consists of blog posts, tweets and social media texts written in both English and Spanish as well as hotel reviews in English. We have considered only the subset which contains blog posts. All bloggers included in the corpus fall into one of these age groups: [18-24], [25-34], [35-49], [50-64], [65-xx]. The corpus incorporates a total of 2278 posts, 148 authors or on an average 15 blogs per author.

We split ourselves into two groups, one for extracting features from the Blog Authorship Corpus (Pranay, Shubham & Soham) and the other from the PAN ’14 Corpus (Aayush, Aseem & Bhushan).

2.2 Data Cleaning & Extraction

- **Blog Authorship Corpus**
  The corpus contains 19,320 XML files, each pertaining to a particular author, identified by the unique filenames. Each XML file contains date when the blog was posted followed by the post itself. All the HTML links in the post are replaced by a unique tag *’urlLink’* to mark their presence. We cleaned the data by discarding empty blog posts and ignoring posts which contain only HTML links and no text. We then exported this refined data to a JSON file, on which further analysis will be carried out.

- **PAN '14 Corpus**
  This corpus contains 148 XML files, each pertaining to a particular author. Each XML file contains the Author's unique ID and blogs written by the Author. The blog text is present in `CDATA` section. To parse this text, we wrote a regular expression to remove the HTML tags, translated HTML entities like `'&amp;'`, `'&ldquo;'` to their usual textual counter-parts like '&',"". We then dumped this data as to a JSON file, on which further analysis will be carried out.

3. APPROACH

After obtaining the JSON files containing refined data from the both the corpora, we now start extracting features from this data-set. We shall be focusing our attention towards building two kinds of classifiers — 1) Binary classifier for classification of gender and 2) Multi-label classifier for classification of age into predefined class labels. Later, we will also consider the possibility of predicting the age by fitting regression models, by working under the assumptions that — 1) There are enough data points for the model to fit accurately and 2) Age behaves like a continuous variable.

3.1 Exhaustive Feature Set

Different people tend to write differently. These differences occur due to variations in the topics of interest and style of writing like word choices and grammar rules. For example, females tend to write more about wedding styles and male tends to write more about technology and politics. Further females use more adverbs and adjectives while writing compared to males. We considered these differences in the writing styles and content of male and female bloggers of different ages. Overall we considered three different types of features that are useful for distinguishing between different categories — content-based, style-based and semantic features. These are enumerated below:

- **Content-based Features**

  Male and female authors tend to speak about different topics, so they will use different words. Thus content based features are important to distin- guish between male and female bloggers

  viz. # of HTML links in the blog, # of named entities used, # of non-word errors, # of discourse relations within the text, # of quotations used in the text, # of references to past or future within the text, # of facts & figures used, # of times opinions are expressed, overall sentiment score of the blog, # of words having character flooding (like 'hellooooo').

- **Style-based Features**

  Features we used include distribution of POS tags, distribution of punctuation tag,s readability measure of the blog (SMOG, Flesch, Gunning Fog etc.), # of co-references (usage of pronouns), average sentence length, usage of figures of speech by the author (like metaphor, alliteration).

- **Semantic Features**

  We shall find a set of topics authors, belonging to a particular gender or age group, blog about. We shall also use Wikipedia concepts and category information to represent the document (blog) [2]. These together will be used as semantic features, as explained in the following section.

These features will be extracted for both the corpora.

3.2 Extracting Feature Vectors

Methodology to extract feature vectors from the Blog Authorship Corpus is described below:

- Content-based Features:
  1) # of HTML links: count the occurrence of *'urlLink'* in the body of the blog.
  2) # of named entities: Using Stanford NLTK APIs to tokenize blog text into sentences, perform POS Tagging and then extract named entities (NE) from the tagged sentences.
  3) # of non-word errors: Using Stanford NLTK's word corpus `nltk.corpus.words.words()` to keep a count of non-word errors in the blog.
  4) # of discourse relations within the text: Using a Java-based end-to-end PDTB-styled Discourse Parser to identify implicit & explicit discourse relations and keeping a count of each of them.
  5) # of quotations used in the text: Checking for occurrences of """ in the blog text.
  6) # of references to past or future: Checking for occurrences of the following set of words and phrases - ['years ago','years from now','in the past','in future','once upon a time','^\d{4}$']. The last entry is a regex for detecting reference to a year.
  7) # of facts & opinions used: Using SentiWordNet to assign score to each blog text; if this score is below 0.1, it is reported as 'fact', 'opinion' otherwise.
  8) Overall Sentiment Score: Review of the PAN '14 Author Profiling Shared Task reveals that incorporating the overall sentiment score of the blog text yields no improvement in accuracy of the classifier. Hence, we shall not consider this feature in this study.

- Style-based Features:
  1) Distribution of POS tags: We use Stanford POS tagger to tag the sentences and collect total counts of all the POS tags for different age buckets & gender. This count is then modeled as a random variable and a CDF is prepared which reports what is the probability that this random variable (say X) will be found to have a value less than or equal to the argument (say x).

  $$F_X(x) = P(X \leq x) \tag{1}$$

  This CDF is prepared for all the age-buckets and both the gender separately.
  2) Readability measure of the blog: We use Stanford NLTK's `punkt` module for assigning readability scores according to different metrics like ARI, Gunning Fog Index, SMOG Index etc. Similar to distribution of POS tags, we shall prepare CDF for the readability score as well.
  3) Usage of Pronouns: Keeping a count of pronouns, used for referencing rather than directly nouns, for all the age-buckets and both the gender.
  4) Average Sentence Length: We keep a average sentence length for the blogs authored by different age buckets and sexes.
  5) Usage of Figures of Speech: To be decided.

- Semantic Features:
  1) LSA to identify topics a person usually blogs about: We use LSA technique to analyze relationships between a set of blogs written by a given author and the terms they contain by producing a set of concepts related to the blogs and terms. We will also try using simple Tf-Idf scores to identify topical words from the blogs authored by a single person, in order to get a set of topics he/she usually blogs about.
  2) Wikipedia Categorization: We shall use Wikipedia concepts and category information to represent the document (blog) [2]

3.3 Feature Subset Selection (FS)

Dimensionality reduction (DR) and Feature Subset Selection (FS) are two techniques for reducing the attribute space of a feature set. The main idea of FS is to remove redundant or irrelevant features from the data set as they can lead to a reduction of the classification accuracy and to an unnecessary increase of computational cost. The advantage of FS is that no information about the importance of single features is lost. For now, we will focus our attention on using FS over DR because DR can decrease the size of the attribute space strikingly. Another important disadvantage of DR is the fact that the linear combinations of the original features are usually not interpretable and the information about how much an original attribute contributes is often lost [7] .If possible, we shall also try using DR technique (PCA) to our feature set and evaluate the improvement in the accuracy of the resulting classifier, if any.
There are three types of feature subset selection approaches:

- Filters:
  Filters are classifier agnostic pre-selection methods which are independent of the later applied machine learning algorithm. . Besides some statistical filtering methods like Fisher score or Pearson correlation, information gain is often used to find out how well each single feature separates the given data set.
  The overall entropy I of a given dataset S is defined as:

  $$I(S) := -\sum_{i=1}^{C} p_i log_2 p_i \tag{2}$$

  where C denotes the total number of classes and p_i the portion of instances that belong to class i. The reduction in entropy or the information gain is computed for each attribute according to:

  $$IG(S, A) = I(S) - \sum_{v \in A} \frac{|S_{A,v}|}{|S|} I(S_{A,v}) \tag{3}$$

  where v is a value of A and $S_{A,v}$ is the set of instances where A has value v.

- Wrappers:
  Wrappers are feedback methods which incorporate the ML algorithm in the FS process, i.e. they rely on the performance of a specific classifier to evaluate the quality of a set of features. Wrapper methods search through the space of feature subsets and calculate the estimated accuracy of a single learning algorithm for each feature that can be added to or removed from the feature subset.

We shall focus our attention towards using Filters to perform FS in this study.

3.4 Building Classifiers & Evaluation

Having a reduced set of features, we shall now try building all types of classifiers available today — Naive Bayes, SVMs, Logistic Regression, k-NN, Decision Trees etc., and evaluate their performance according to the following metrics:

- Precision

  $$P = \frac{TP}{TP + FP} \tag{4}$$

  The percentage of positive predictions that are correct.

- Recall

  $$R = \frac{TP}{TP + FN} \tag{5}$$

  The percentage of positive labeled instances that were predicted as positive.

- Specificity

  $$S = \frac{TN}{TN + FP} \tag{6}$$

  The percentage of negative labeled instances that were predicted as negative.

- Accuracy

  $$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

  The percentage of predictions that are correct.

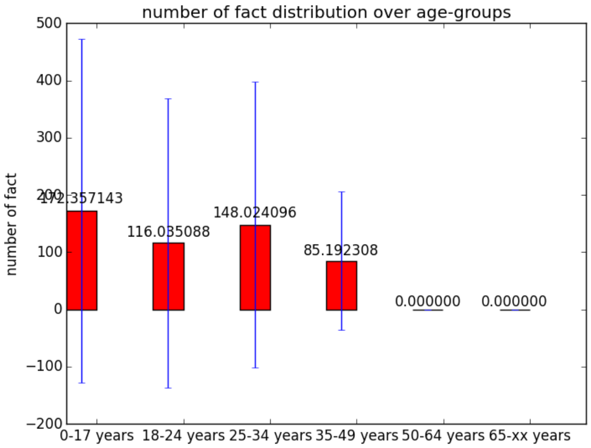
  If time permits, we shall also explore the possibility of modeling the prediction of age as a regression problem and try to improve the baseline established by [3].

4. RESULTS SO FAR

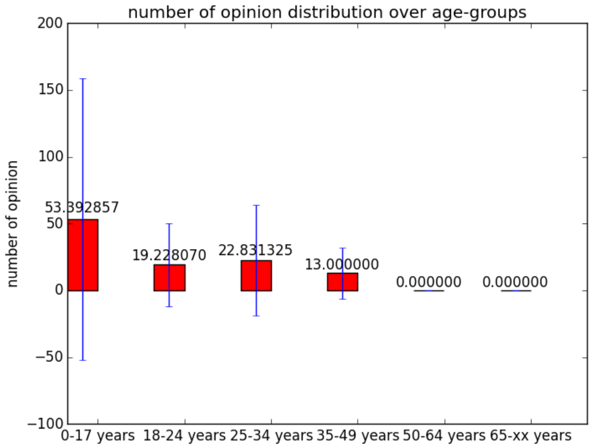
We have looked into influence of individual features in prediction of the age and gender. And because of high variance in these plots, it is clear that individual features, which we think are important distinguishing elements, don't play much role in the personality detection. We need to consider a combination of these in order to get something meaningful. We will next try to learn a classifier, that learns the weight given to each feature. This way we would be able to consider a combination of features and we'll also learn how much important a feature is in the required task.

4.1 Blog Authorship Corpus

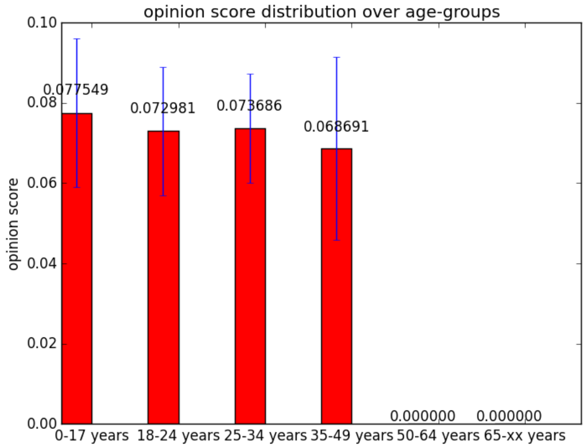
Here goes the description about the plots obtained from the Blog Authorship Corpus.



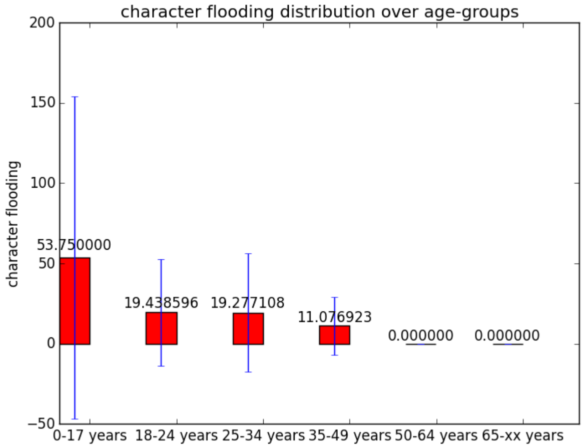
(a) # of Facts vs Age



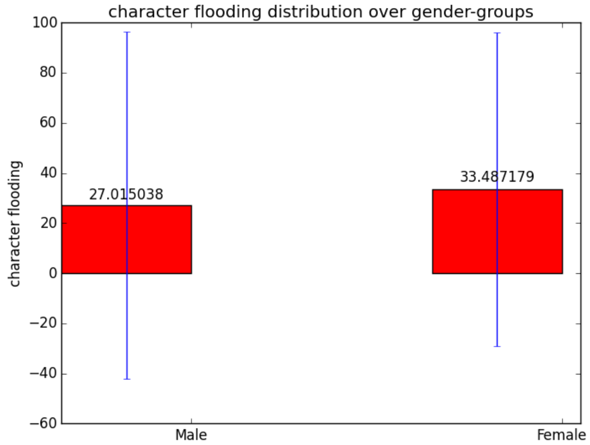
(b) # of Opinions vs Age



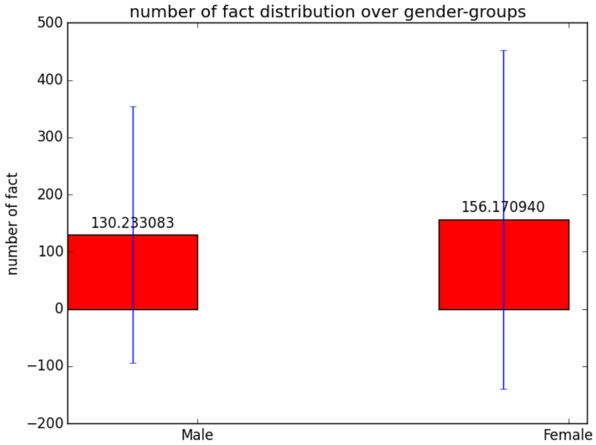
(a) Blog Opinion Score vs Age



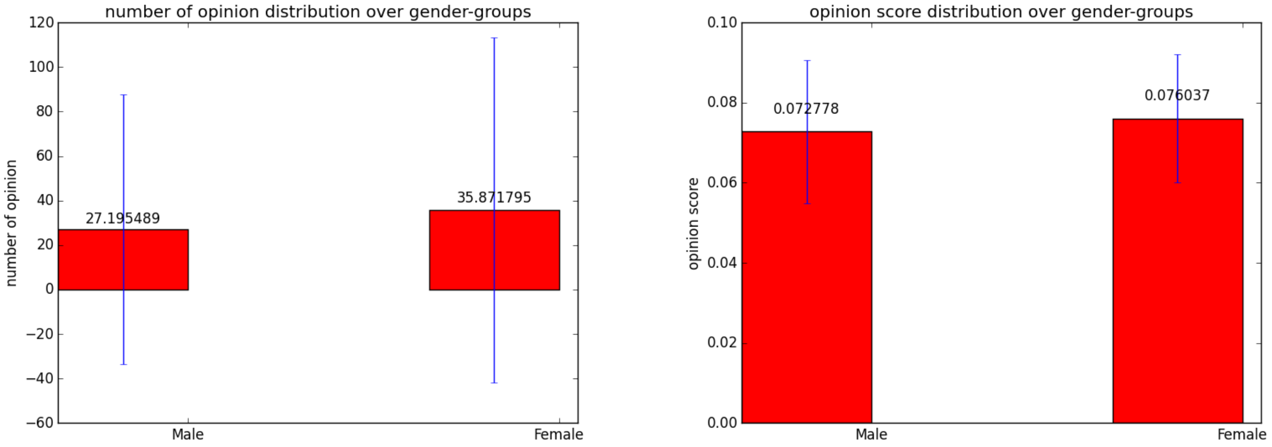
(b) # of instances of Character Flooding vs age



(a) # of instances of Character Flooding vs Gender

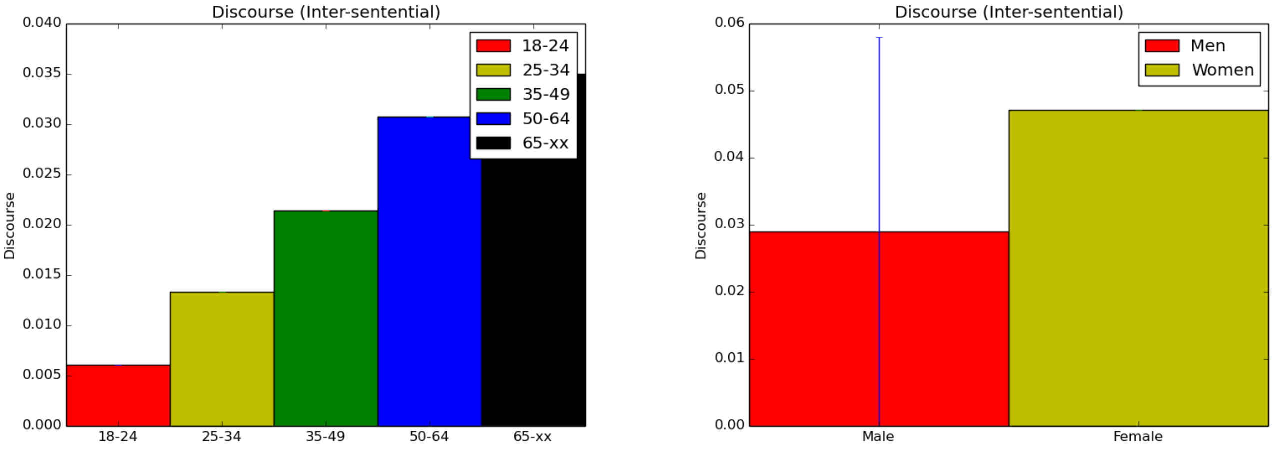


(b) # of Facts vs Gender



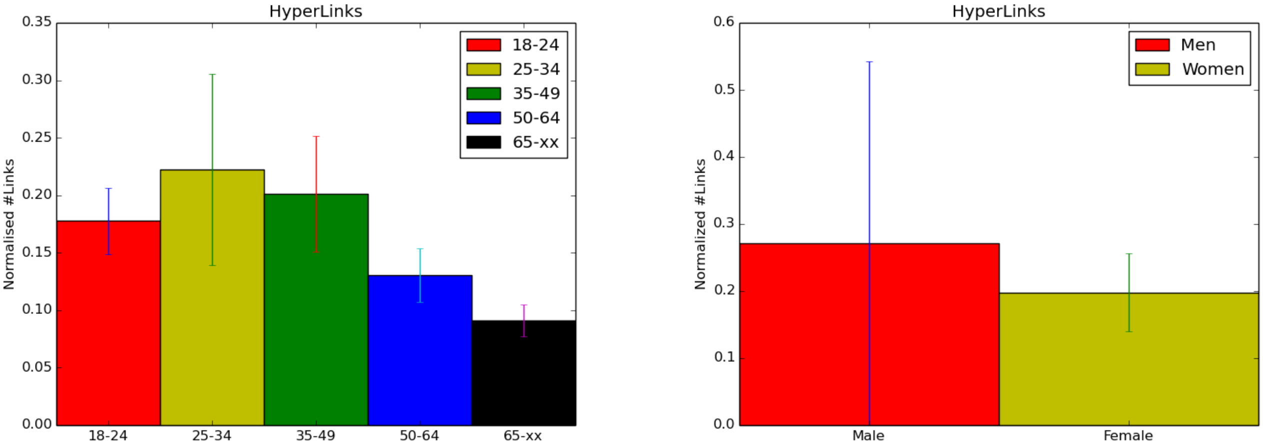
(a) # of Opinions vs Gender

(b) Blog Opinion Score vs Gender



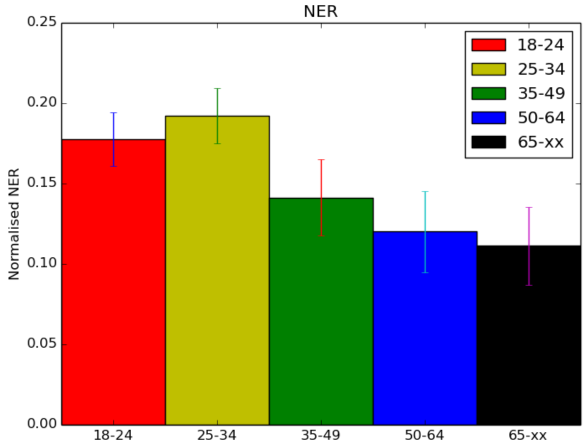
(a) # of Inter-sentential Discourse Relations vs Age

(b) # of Inter-sentential Discourse Relations vs Gender

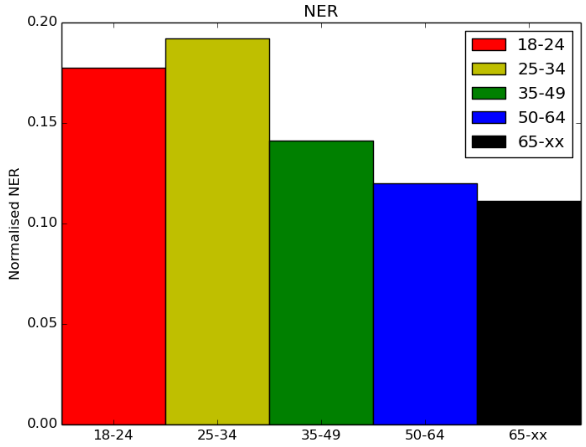


(a) Average # of HTML Links vs Age

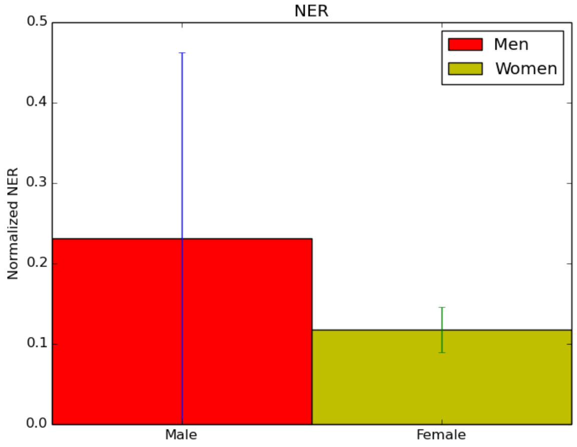
(b) Average # of HTML Links vs Gender



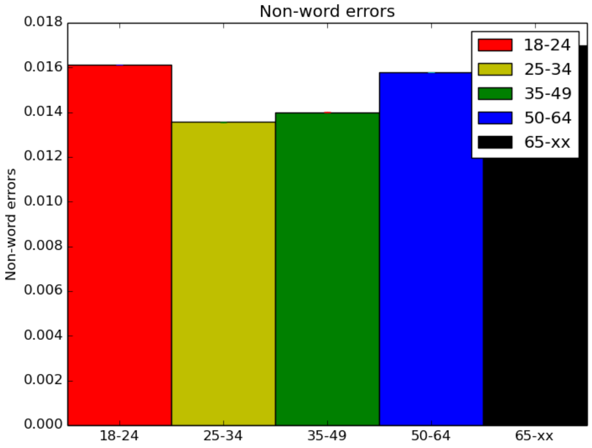
(a) Average # of Named Entities used vs Age



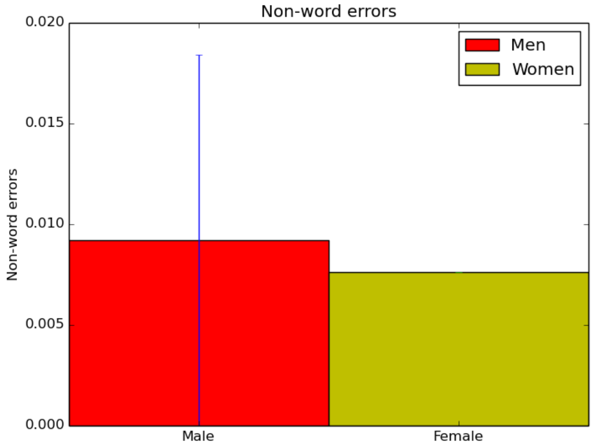
(b) Average # of Named Entities used vs Age



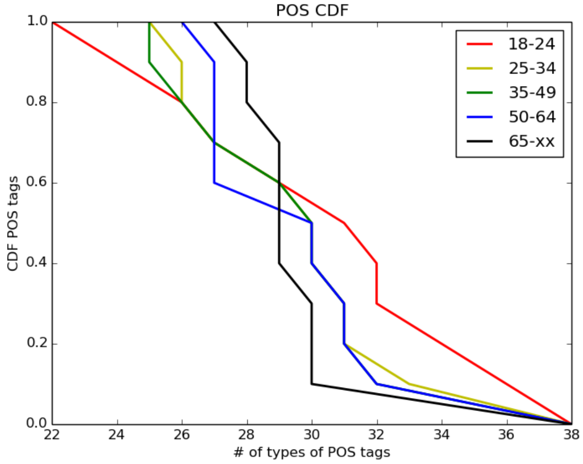
(a) Average # of Named Entities used vs Gender



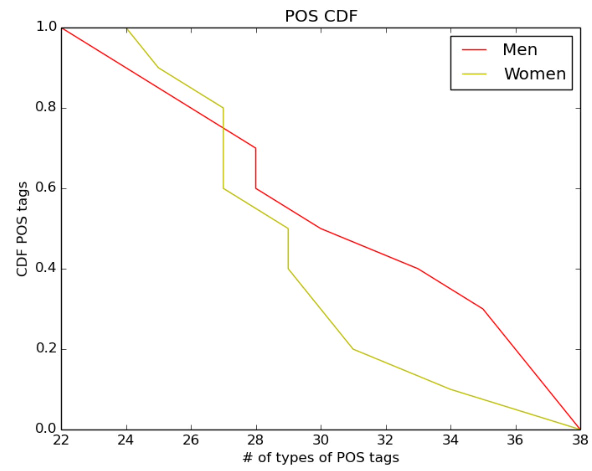
(b) # of Non-word Errors vs Age



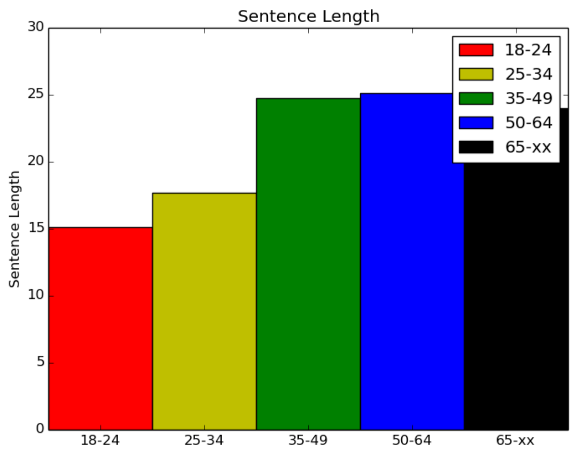
(a) # of Non-word Errors vs Gender



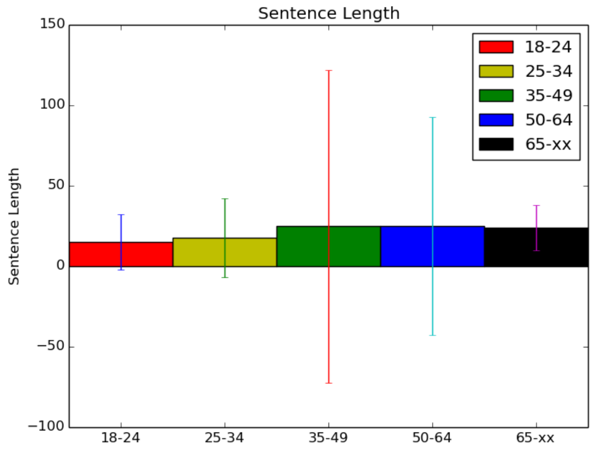
(b) Cumulative Distribution of POS tags vs Age



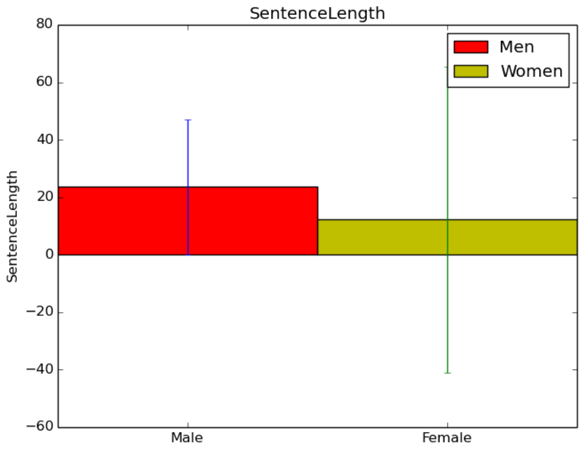
(a) Cumulative Distribution of POS tags vs Gender



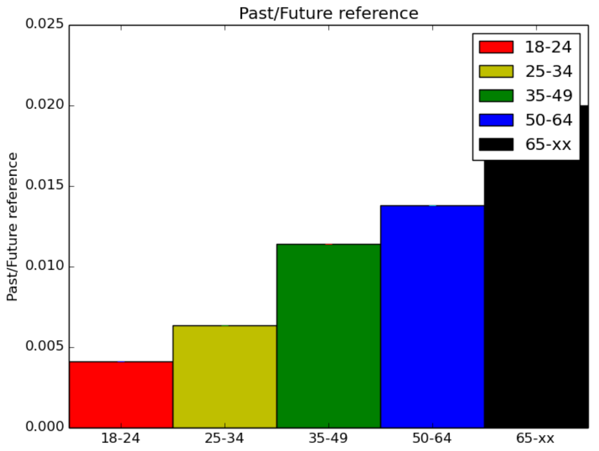
(b) Average Sentence Length vs Age



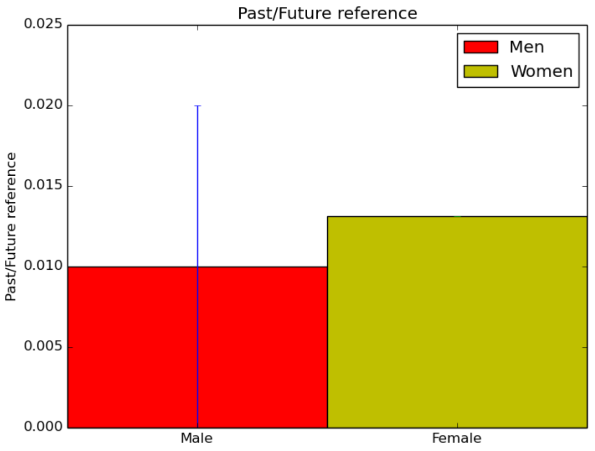
(a) Average Sentence Length vs Age



(b) Average Sentence Length



(a) Average # of Future/Past time references vs Age



(b) Average # of Future/Past time references vs Gender

For Age Bucket Classification:

- # of facts vs Age — contains average # of factual blogs vs age buckets
- # of opinions vs Age — contains average # of opinion blogs vs age buckets
- Blog opinion score vs Age — contains average blog opinion score vs age buckets
- # of instances of character flooding vs Age — contains average # of instances of character flooding (like 'hellooo') vs age buckets
- Average # of inter-sentential discourse relations vs Age: contains average normalized # of inter-sentential discourse relations (X) vs age buckets

$$X = \frac{\text{\#of inter-sentential discourse relations in a blog}}{\text{\#of sentences in that blog}} \tag{8}$$

- Average # of HTML links vs Age — contains average normalized # of HTML links (X) vs age buckets

$$X = \frac{\text{\#of HTML links in a blog}}{\text{\#of sentences in that blog}} \tag{9}$$

- Average # of named entities used vs Age— contains average normalized # of named entities (X) vs age buckets

$$X = \frac{\text{\#of named entities in a blog}}{\text{\#of sentences in that blog}} \tag{10}$$

- Average # of non-word errors vs Age— contains average normalized # of non-word errors (X) vs age buckets

$$X = \frac{\text{\#of non-word errors in a blog}}{\text{\#of words in that blog}} \tag{11}$$

- Distribution of POS tags vs Age — contains Cumulative Distribution of usage of POS tags vs age buckets
- Average sentence length vs Age: contains average sentence length vs age buckets
- Average # of future/past time references vs Age — contains average normalized # of future/past time references (X) vs age buckets

$$X = \frac{\text{\#of future/past time references in a blog}}{\text{\#of sentences in that blog}} \tag{12}$$

For Gender Classification: Similar plots, as for Age Bucket Classification.

4.2 PAN'14 Corpus

Here is the description about plots obtained from PAN'14 Corpus.

For Age Bucket Classification:

- Average # of HTML links vs Age — contains average normalized # of HTML links (X) vs age buckets

$$X = \frac{\text{\#of HTML links in a blog}}{\text{\#of sentences in that blog}} \tag{13}$$

- Average # of named entities used vs Age— contains average normalized # of named entities (X) vs age buckets

$$X = \frac{\text{\#of named entities in a blog}}{\text{\#of sentences in that blog}} \tag{14}$$

- Distribution of POS tags vs Age — contains Cumulative Distribution of usage of POS tags vs age buckets
- Average sentence length vs Age: contains average sentence length vs age buckets
- Average # of quotations vs Age— contains average normalized # of quotations (X) vs age buckets

$$X = \frac{\text{\#of quotations in a blog}}{\text{\#of sentences in that blog}} \tag{15}$$

- Various Readability Scores vs Age — We computed several readability metrics *[3.2]* like Automated Readability Index(ARI), Flesch Reading Ease, Flesch Kincaid Grade Level, Gunning Fog Index, SMOG Index, Coleman Liau Index, LIX and RIX for the blogs; It is found that as age increases, readability of the text written by that person decreases. This aligns with our intuition. For gender, blogs written by males tend to have less readability that than female authors.

For Gender Classification: Similar plots, as for Age Bucket Classification.

5. FUTURE WORK-PLAN

5.1 Feature Selection

We shall try using both the techniques for Feature Selection namely — Feature Subset Selection (FS) using Filters/Wrappers and Dimensionality Reduction (DR) using PCA. The primary focus will be on using Filters for Feature Selection task using Information Gain (IG), as quoted earlier in *[3.3]*.

5.2 Incorporate Semantic Features

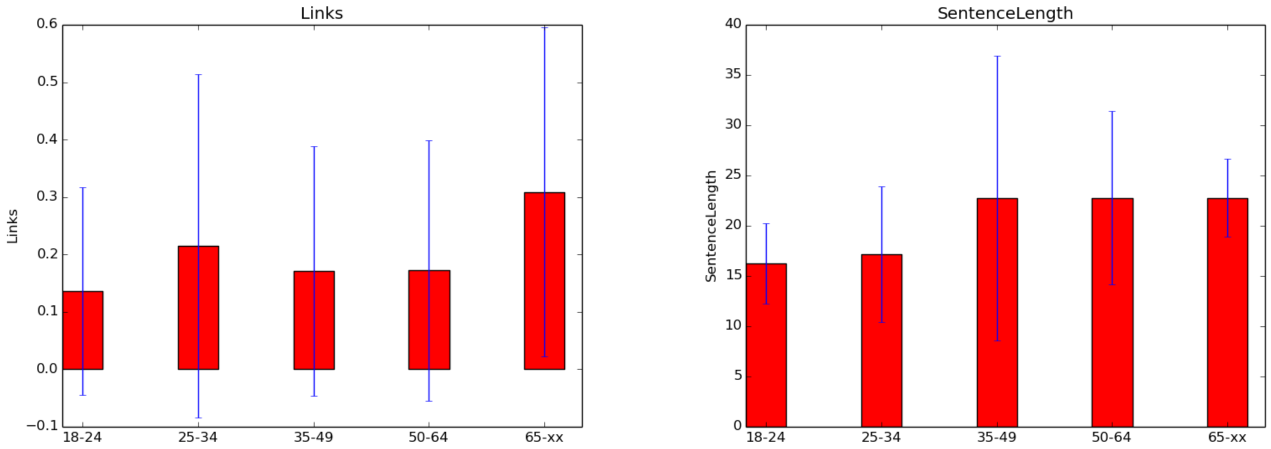
We expect semantic-based features to add to the performance of the classifier because of the following reasons: 1) Content-based or Style-based features alone do not consider the semantic relation between words. 2) These do not handle polysemy. [**?**] shows that classifiers learned using semantic features based on Wikipedia Category Information achieve significantly better accuracy compared to the state-of-the-art methods. Hence, we aim to cover this and other semantic features in our classifier as well.

5.3 Deep Learning

Deep Learning techniques have proven themselves to have an astonishing performance over the conventional methods. Using Word2Vec, we shall try incorporating deep learning approach for out task and evaluate its performance vs the state-of-the-art methods.

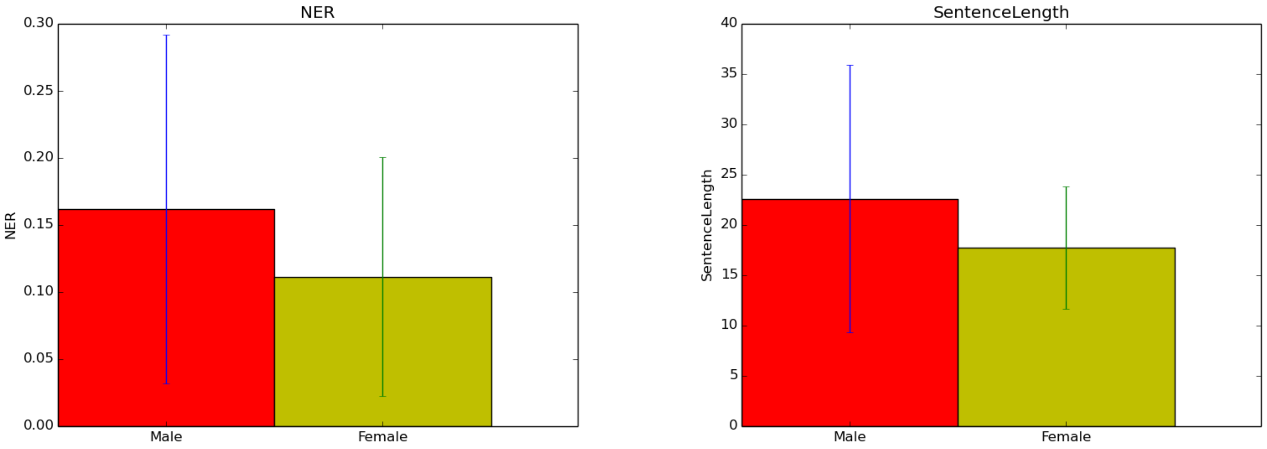
5.4 Regression Modeling

With reference to the regression model discussed in this study — [3], we will compare the performance of the regression approach for modeling the age of the author vs our classification approach to identify the age bucket.



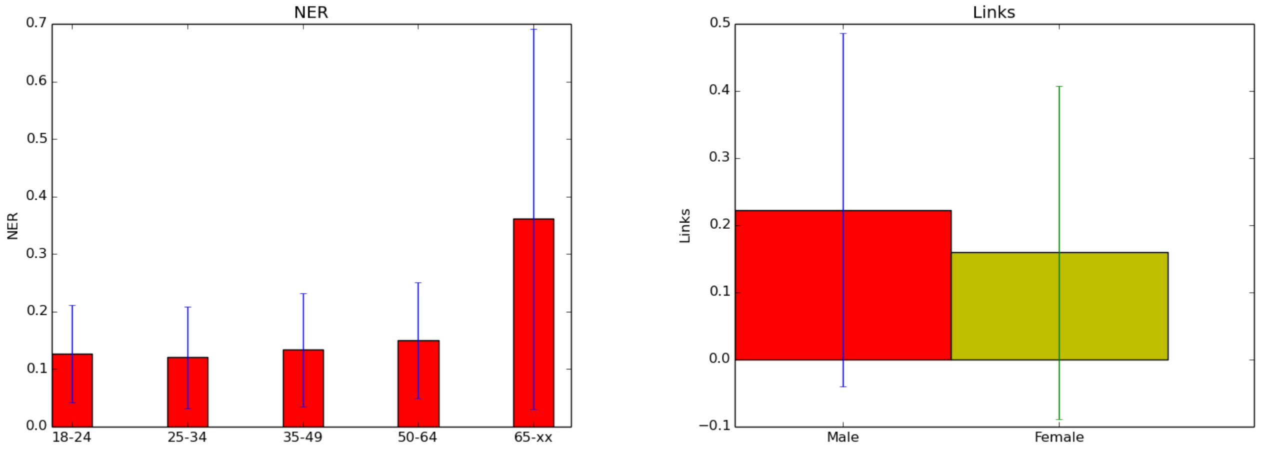
(a) # of HTML Links vs Age

(b) Average Sentence Length vs Age



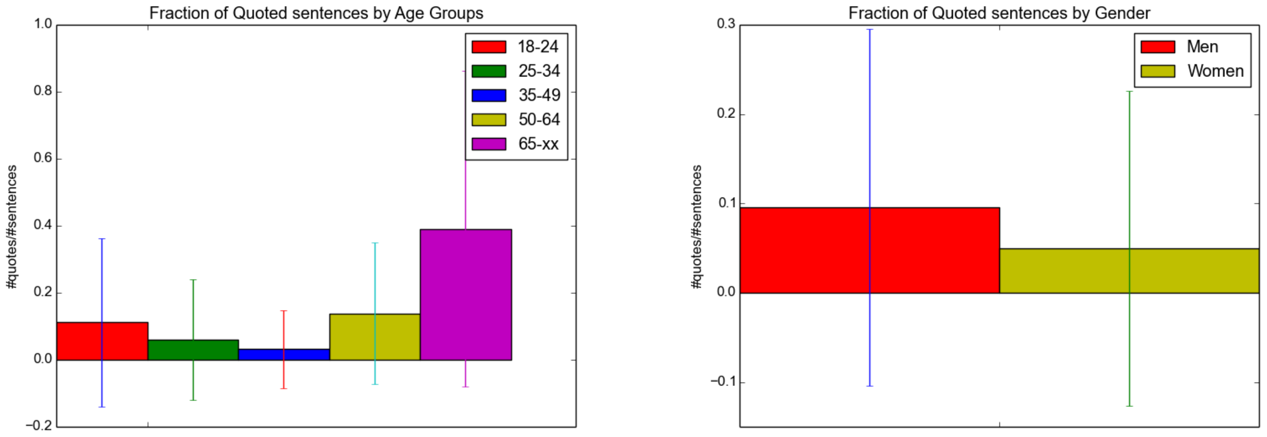
(a) # of Named Entities used vs Gender

(b) Average Sentence Length vs Gender



(a) # of Named Entities used vs Age

(b) # of HTML Links vs Gender



(a) # of Quotations vs Age

(b) # of Quotations vs Gender

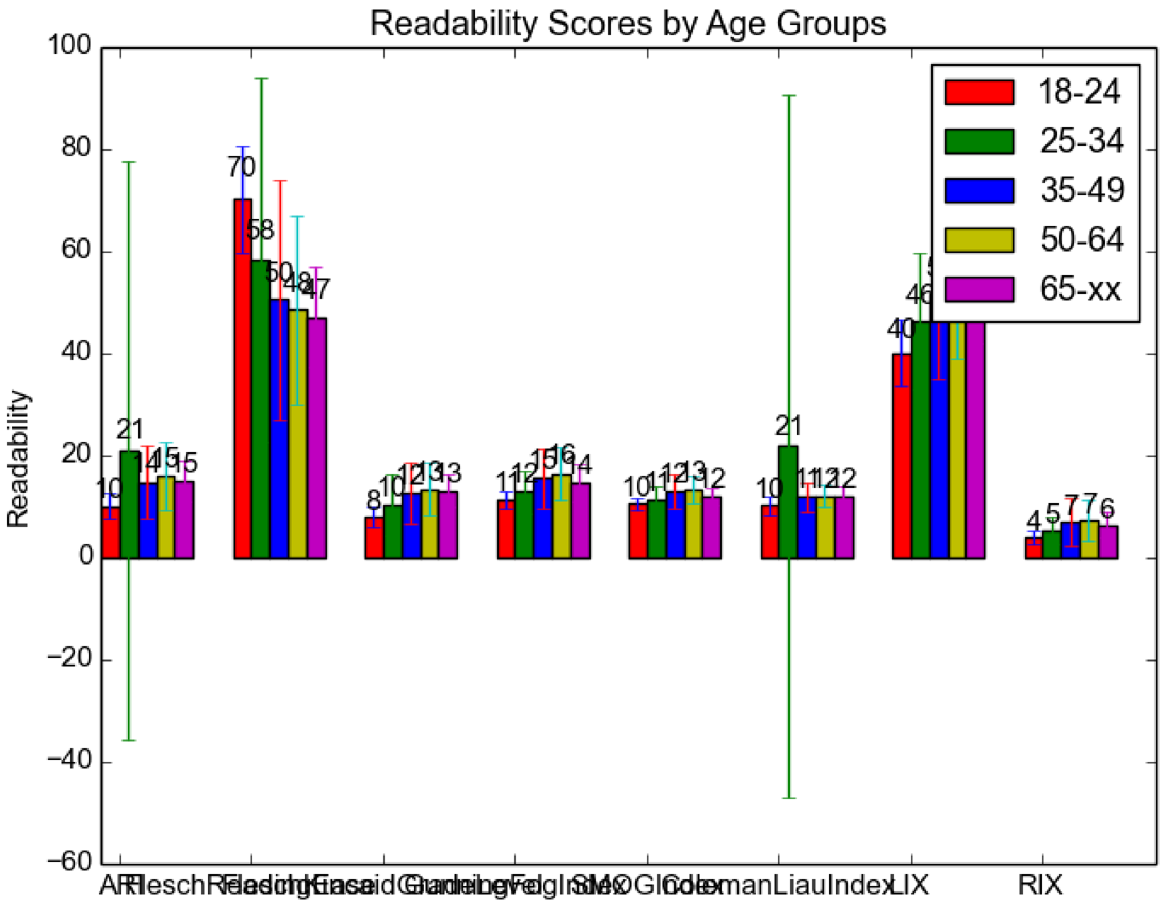


Figure 1: Readability Scores vs Age

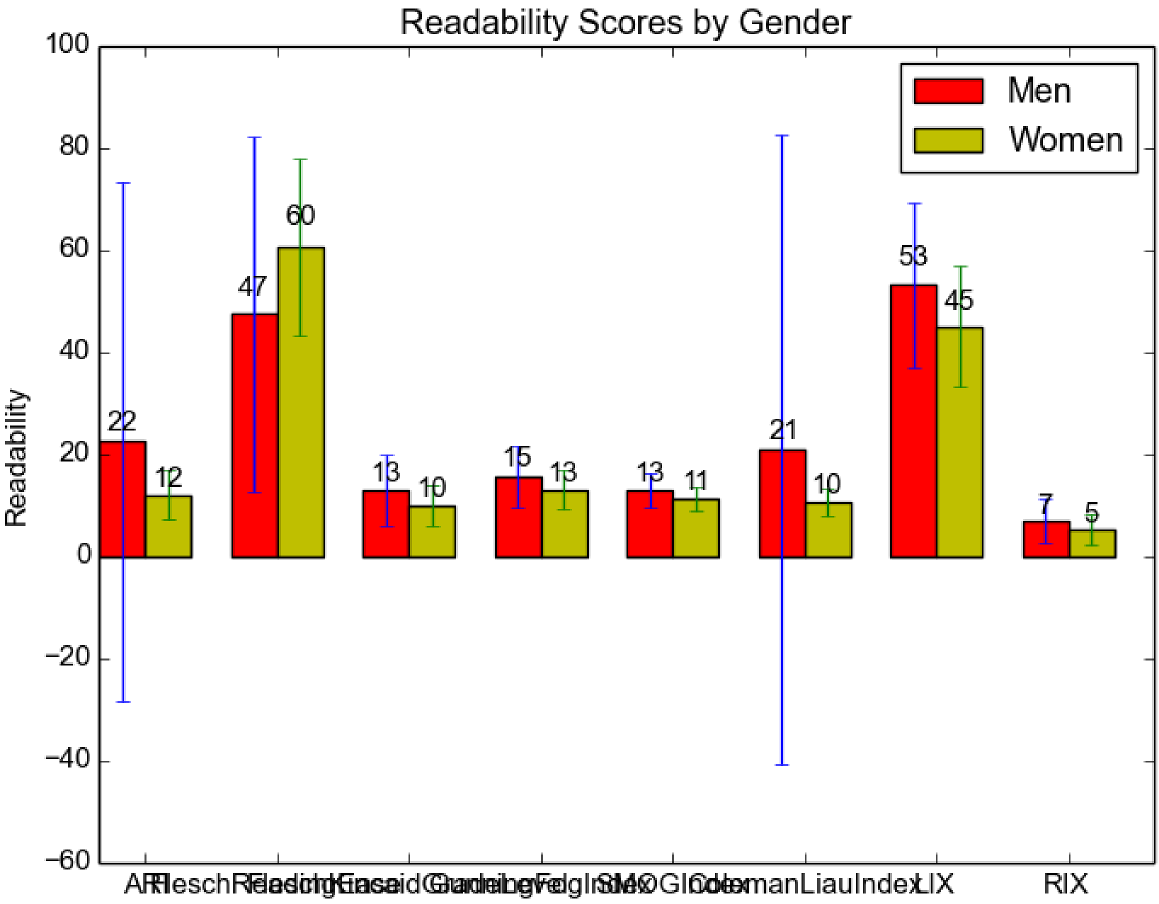


Figure 2: Readability Scores vs Gender

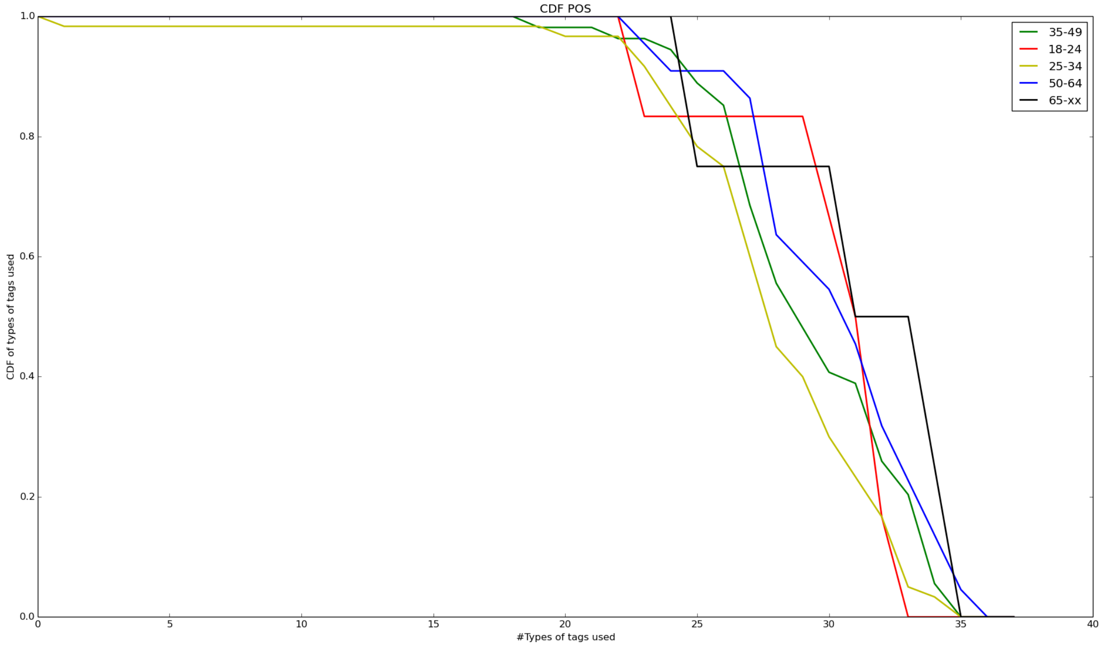


Figure 3: Cumulative Distribution of POS tags vs Gender

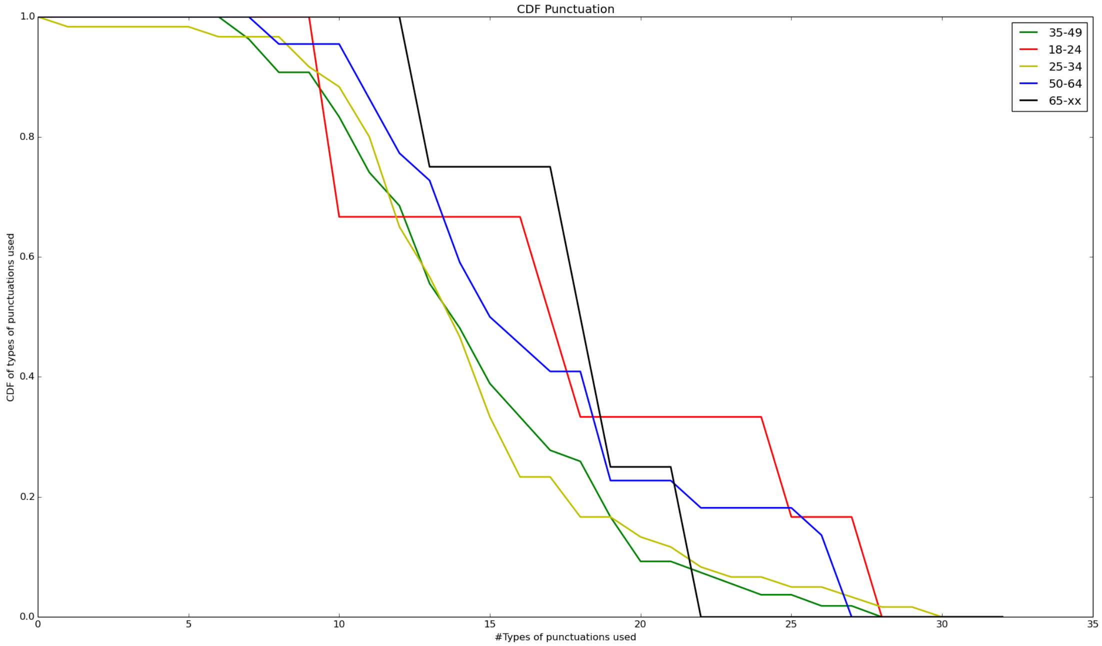


Figure 4: Cumulative Distribution of POS tags vs Age

6. PERFORMANCE EVALUATION

For the Blog Authorship Corpus, we plan to use k-Fold Cross Validation for evaluation of classifiers, as quoted in *[3,4]*. For the PAN'14 Corpus, we have a test data-set (corresponding to PAN' 15) which is unlabeled and the tagged data-set will be released in recent future. If we don't get access to that soon, we will go with k-Fold Cross Validation on the PAN' 14 data.

7. REFERENCES

[1] James Marquardt et al. [*Age and Gender Identication in Social Media*]. Annalen der Physik, 322(10):891-921, 1905.

[2] K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma *Exploiting Wikipedia Categorization for Predicting Age and Gender of Blog Authors.* Notebook for PAN at CLEF 2013.

[3] Dong Nguyen, Noah A. Smith, and Carolyn P. Rose *Author Age Prediction from Text using Linear Regression.* In Proceedings of the ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LATECH 2011), Portland, OR, June 2011.

[4] Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. *Overview of the Author Profiling Task at PAN 2013.* Proceedings of PAN at CLEF 2013.

[5] S. Argamon, M. Koppel, J. Pennebaker and J. Schler (2009) *Automatically profiling the author of an anonymous text.* Communications of the ACM 52 (2): 119-123.

[6] J. Schler, Moshe Koppel, S. Argamon and J. Pennebaker (2006) *Effects of Age and Gender on Blogging* in Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, March 2006.

[7] Andreas Janecek, Wilfried Gansterer, Michael Demel, Gerhard Ecker *On the Relationship Between Feature Selection and Classification Accuracy* JMLR W&P 4:90-105, 2008.