Literature Survey

Understanding the variation of language use among Bengali users of based on their age and gender.

Srija Mukhopadhyay (2021114002) - *srojo*

Stylometric Analysis of Bloggers' Age and Gender

Goswami, S., Sarkar, S., Rustagi, M. 2009. Stylometric Analysis of Bloggers' Age and Gender

This paper uses data taken from blogs to analyse and predict the gender and age of the writers. They have commented on how bloggers generally express their thoughts in an informal, unreserved and unorganized manner through blogs, which is likely to make the data that is presented more candid and something that is more likely to have distinguishable features based on gender and age.

The data they used was the blog corpus available in the website of Prof Moshe Koppel, which had blogs with author-provided indication of both gender and age.

The features that were used for the detection were sentence length and non dictionary words.

Non dictionary words are words which are understandable and commonly used by online communities, and they include things such as slangs, smiley, out of dictionary words, chat abbreviations, etc. The named entities are also non dictionary words. There are also words which are intentionally misspelled, repeated, extended or shortened to have a different effect on the reader, express emotion or save the time of blogging. All these words and the frequency of use of such words contribute to the features in stylometric.

It was observed that teenagers generally use more non dictionary words than the adults. In addition to that, for bloggers of each gender, there is a clear distinction between usages of a few slangs.

Average sentence length as a feature is pretty self explanatory, however it was realised that it alone as a single feature can't make a good classifier.

It was observed that teenagers used smaller sentences compared to adult bloggers. However, owing to the fact that there is a dearth in data, the authors couldn't conclude if the average sentence length increased with age.

The authors tried experimenting with different models for the features enlisted, along with another additional feature which was the set of 30 words that Koppel listed as a distinguishing feature for gender and age, referred here as

'content words' and were seen to have an extreme variation in usage across gender and age groups.

For the feature of non dictionary words, Naïve Bayes Classifier yielded an accuracy of 77.39 % for gender based classification and 89.68 % accuracy for the age group classification between 10s and 30s age. Augmenting the feature list with the content words and non dictionary words yielded an accuracy of 80.32% for age and 89.18% for gender classifiers. Adding average sentence as a feature increased that accuracy to 80.38% and 89.30% respectively.

Evaluation and Sociolinguistic Analysis of Text Features for Gender and Age Identification

Simaki, V. Mporas, I. Megalooikonomou, V. 2016. Evaluation and Sociolinguistic Analysis of Text Features for Gender and Age Identification

In this paper, the authors were interested on the demographic information of users (gender and age) and how that can be derived from linguistic clues only. The characteristics used in a wide set of text mining tasks were investigated and collected and their efficacy for the age and gender classification tasks were evaluated.

In addition to that, the most significant clues have been associated to existing sociolinguistic markers and conclusions about the nature of important features of gendered and age differentiated linguistic choices have been derived.

Another important feature that was analysed was the question if the knowledge of a social variable (gender or age) is helpful through an identification process of the other variable. In other words, could the gender be calculated as a classification feature in an age identification task and vice versa, is age a differential characteristic of gender identification.

The authors attempted an interdisciplinary study, combining the sociolinguistic knowledge on gendered and age linguisitc variation to the existing text mining techniques for the author's profile exploitation.

These features were extracted and ranked according to their contribution in the author's gender and age detection.

The corpus used for this study was the "Blog Authorship Corpus", primary due to its annotation with both gender and age information. A second reason that attracted the researcher' interest was the text type of the posts: informal and spontaneous text samples produced by social media users.

The features that were used for the process were

- The statistical features, forming a feature vector equal to 30 - The POS-tags features, forming a feature vector equal to 9 - The content-based features, forming a feature vector equal to 3 - The gender feature - which is used for the identification of the age variable - The age feature - which is used for identification of the gender variable

For analysing how important each feature is, the Relief feature selection

algorithm was used (the updated ReliefF algorithm was used as proposed by Koronenko (1994), which improves the reliability of the probability approximation and is robust to incomplete data and generalized to multi-class problems.

The dataset was processed by the ReliefF algorithm, implemented using the WEKA machine learning toolkit and feature ranking scores were estimated.

It was found that for the age estimation, 18 out of 43 features proved to the significant with gender being placed 14th.

On the other hand, for the gender estimation, 37 our of 43 features proved to the significant with age being placed in the 4th position.

The feature set used proved to be more effective in the gender identification task, highlighting more informative features, than during the age identification task.

They commented that more extensive study of the important features that are common in both tasks, may lead to a grouping of these characteristics in terms of the linguistic level of analysis they are located: morphological level, lexical level, syntactic level and context level.

The authors did comment that the morphological features need to be more investigated and analyzed in a set of different tasks for the gender and age identification, in order to enable the outcome of more specific conclusions. Furthermore, it has to be compared the values of these features given a different variable (gender/age).

On the other hand, for the syntactic features, they said that they could have an idea about their informatively, based on research in the fields of sociolinguistics and the automatic gender and age identification: the use of pronouns and adverbs are strongly related to the gender and age linguistic variation as age and gendered preferential choices. The new clues though, of the use of articles and prepositions should be in future work be more investigated and analysed as the authors commented.

The statistical features were found to be more numerous than the other features.

Concerning the age detection task, statistical features based on character calculations proved to be very important and and can be associated to existing sociolinguistic knowledge on age linguistic variation. The word length or the short words have been observed as markers distinguishing the linguistic use that teens and adults make.

On the other hand, POS features were found to be critical in distinguishing the gendered linguistic choices of a user.

In addition to that, 9 over the 18 most informative clues are common in both tasks and the knowledge of gender and age is of great importance as feature for the corresponding investigation.

The authors did finally conclude by saying that in order to confirm the sociolinguistic indications concluded by the present study to standard theories.

Effects of Age and Gender on Blogging

Schler, J., Koppel, M. Argamon, S., Pennebaker, J. 2006 Effects of Age and Gender on Blogging

The main aim of the study was to use the rich data available from blogs (because of the no restriction on choice of topic) to answer the following two questions - How do content and writing style vary between male and female bloggers and among bloggers of different ages?

For the purpose of this study, the authors reviewed all blogs accessible from blogger.com one day in August 2004, downloading each one that included author-provided indication of gender and at least 200 appearances of common English words. The full corpus thus obtained included over 71000 blogs.

To prevent bias, the authors created a sub-corpus consisting of an equal number of male and female blogs in each age group, by randomly discarding surplus documents in the larger category.

They used two distinguishing features which were style related and content related.

For the style related features, the things considered were parts of speech, function words and blog specific features such as "blog words" and hyperlinks.

On the other hand, the content related features were simpler content words as well as special classes of words taken from the handcrafted LIWC categories.

For the style based features, it was noted that for each age bracket, female bloggers use more pronouns and assent/negation words while male bloggers use more articles and prepositions. Also, female bloggers were seen to use blog words far more than do male bloggers, while male bloggers use more hyperlinks.

Another interesting thing to note was the fact that prepositions and articles, which are used more frequently by male bloggers are used with increasing frequency by all bloggers as they get older.

Conversely, pronouns, assent/negation words and blog words, which are used more frequently by female bloggers, are used with decreasing frequency as bloggers get older.

In short, the very same features that distinguish between male and female blogging style also distinguish between older and younger blogging style.

In the case of content based features, the differences suggested a pattern of more personal writing by female bloggers than male bloggers. Male blogging is characterized by far more references to politics and technology.

It is seen that as age progresses, teenage concern with friends and mood swings gives way to the indulgences of college life and then eventually to marriage, its attendant financial concerns and an interest in politics.

It should be noted that the full list word frequencies broken down by age indicates that almost all words increase or decrease monotonically with age, and words that increase monotonically with age are "male" words while those which decrease with age are "female" words.

The researches also carried out a process of automatic author profiling based on their research, where the stylistic features mentioned previously along with

the 1000 unigrams with highest information gain in the training set for the content words.

It needs to be obviously noted that blogs contain other clues to bloggers' age and gender, like blogger provided profile information as well as subtle clues like choice of formatting template and color and use of emoticons and other non lexical features. These information was omitted although it had a quite useful value otherwise.

The learning algorithm used was Multi Class Real Winnow to learn models that classify blogs according to author gender and age respectively and using all the features an accuracy of 80.1% was obtained.

In addition to that, the confusion matrix indicated that using content and style features together, 10s are distinguishable from 30s age-wise, with accuracy of over 96% and distinguishing 10s from 20s is also achievable with accuracy of 87.3% but many 30s are misclassed as 20s, bringing the accuracy down to 76.2%.

The authors finally concluded that male bloggers of all ages write more about politics, technology and money than do their female cohorts. Female bloggers discuss their personal lives – and use more personal writing style – much more than males do. Furthermore, for bloggers of each gender, a clear pattern of differences in content and style over age is apparent. Regardless of gender, writing style grows increasingly "male" with age: pronouns and assent/negation become scarcer, while prepositions and determiners become more frequent. Blog words are a clear hallmark of youth, while the use hyperlinks increases with age. Content also evolves with age in ways that could have been anticipated.

Conclusion and Inferences

Reading the papers gave me an in-depth insight into the methodologies I can use going forward with my project. I do however want to spend some more time reading papers (as indicated in the project proposal) before formulating a proper roadmap for how I want to carry out the entire process.