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

Srija Mukhopadhyay
2021114002
International Institute of Information Technology, Hyderabad
srija.mukhopadhyay@research.iiit.ac.in

Abstract

The paper aims to understand if there is any variation in the use of Bengali among users of different age and gender groups. Such variation has already been observed in English, but there isn't a lot of progress in the same domain when it comes to Bengali and other Indian Languages. The paper has investigated on the same and tried to find conclusive results through the process of extensive research and study.

1 Introduction

The user of any language has a unique way of speaking or communicating. However in addition to that, there are a few features which are common across different social groups.

Usually we study the variation of language across different social groups or categories in Sociolinguistics.

The aim of this project is to find the variation of language use among Bengali users across different age and gender categories.

Only written data was used for the purpose of this project as it is easier to obtain and analyse over a limited period of time.

In addition to that, the initial hypothesis of the project was that there won't be significant variation across the different age and gender categories owing to the fact that the way Bengali is used is mostly uniform, especially in written form.

It usually lacks things such as abbreviations and slang terms which often help in performing a similar language based analysis for languages such as English. The only possible difference that could be thought of was the use of emojis in online text, which is also expected to be sparse and not too commonly used.

Owing to that fact, a mostly uniform language use pattern across different age and gender categories was expected before starting the project, as an initial hypothesis.

2 Previous Work

Before starting the work on the project, a literature survey was conducted, where three papers gave valuable insights into the entire topic and acted as pillars on the basis of which research for this project was conducted and gave direction to the entire course of the project.

One of the papers was "Stylometric Analysis of Bloggers' Age and Gender" which tried to identify age and gender variation on the basis of factors such as sentence length and the use of non-dictionary words.

They also used important words, defined as "content words" in order to classify data and train a model which could use the aforementioned features to identify the gender and age among bloggers.

Another paper, titled "Evaluation and Sociolinguistic Analysis of Text Features for Gender and Age Identification" was also really helpful in highlighting the importance of text mining tasks to highlight differences between the categories of age and gender.

The paper concentrated on different lexical, syntactic and morphological features to find out which features mattered the most for the process of age and gender identification.

In addition to that, they also analysed the importance of age data for the classification and identification of gender and the importance of gender data for the classification and identification of age.

While the paper commented on the scope of exploring about the importance of morphological features, it noticed that POS features

were critical in distinguishing gendered linguistic choices of an user.

Finally, another paper titled "Effects of Age and Gender of Blogging" was pretty helpful as it analysed the kind of words that are present across the different categories and the variation in their presence in the different cases.

They commented on things such as how a teenager is more likely to talk about friends and school and mood swings while later it progresses to talk about things like marriage, financial concerns and politics.

Thus content words like those were concluded to have a pretty decent influence in helping identify the age and gender categories according to that research conducted.

All the above mentioned researches had their work conducted in English. Coming to Bengali, not a lot of work has been conducted in this topic.

The only paper that came remotely close to the topic was "DAAB: Deep Authorship Attribution in Bengali" which worked on created a neural network model for authorship attribution and identification and not much on the variation of language with respect to age and gender.

Thus, one can say that the work conducted in this paper is pretty novel, especially from the view of using Bengali as the language that is studied.

3 Methodology

3.1 Data Collection

For the process of data collection, we used the Bengali Quora website because it allows for free use of the language without any specific set of language rules that need to be followed allowing the writers to write more candid, informal, unreserved and unbiased texts, as well as the liberty that the users get to write on any topic of their choice.

The user base was divided into two gender categories - Male and Female and 3 age categories - 10 to 20, 20 to 30 and 30 to 50. These were the only age categories where data could be found in the website and thus including other age categories were not possible.

The age category of 30 to 50 was clubbed together because there is hardly much variation in patterns of those age categories. (As highlighted by other papers read during the literature review)

The process of data collection was made hard by the following two facts

- Finding profiles of the required age and gender categories was pretty hard. That along with the fact that the ages and gender categories needed to be identified manually made the process even more tedious.
- There are quite a few spam profiles on the website, which are mostly there for advertisements or users trying to get quick attention on the website (find some reference for this as well), as a result those profiles needed to be filtered out to ensure proper data collection

For the process of collecting data, Node.JS was used for web scraping along with Selenium for proper parsing of the website and finding the right amount of data.

The following methodology was implemented for the same

- Going to a specific topic and finding the top 10 writers of that topic. While choosing the different topics, I tried to ensure that a variety of topics were chosen to ensure the absence of any bias in the data. It should also be noted that just because a certain user is the top writer of a given topic does not necessarily mean that all the extracted answers by that writer will belong to that specific topic.
- Identifying the age and gender of the user. This was done manually, with the help of the name shown on the profile and other user details mentioned on the user profile. In addition to that, sometimes the user was looked up on Google to obtain more concrete data about the user.
- Extracting a certain number of answers written by that user. The number of answers extracted varied according to their age and gender to compensate for the lack of data for certain categories.
- Annotating the data extracted as it is saved to a file

The aforementioned process ensured that a pretty extensive corpora is created for the entire

process and also gave valuable insights on the gender and age distribution of authors on the basis of the different topics that were explored. The same is explained in detail in the observations section.

The initial aim was to find at least 250 answers for each age category and at least 500 answers for each gender category along with a proper distribution between the co-variation in the two categories.

Eventually, we ended up getting more data, although it was hard to establish a proper covariance in the distribution but the best efforts were made to do so.

The profiles of 25 unique users were scrapped in order to form the annotated corpora.

The following table portrays the actual amount of answers obtained for each category for the preparation of the corpora.

Category	Users	Answers
Male	12	900
Female	12	900

Table 1: Table to show distribution of data with respect to gender

Category	Users	Answers
10 to 20	5	400
20 to 30	12	720
30 to 50	7	680

Table 2: Table to show distribution of data with respect to age

Once the data was collected and annotated properly, it was put into respective files where it was analysed for the purpose of this report.

3.2 Data Analysis

Once the data was collected, it was put into 5 text files according to the age and gender categories to enable the process of further analysis. Each of the text file had to undergo the same process for analysis.

The initial part was counting the number of emojis used in each of the text files to account for the number of non dictionary words present in text.

The distribution of emojis across the different categories can have been highlighted in the following tables.

Category	Emojis
Male	453
Female	673

Table 3: Table to show emoji distribution of data with respect to gender

Category	Emojis
10 to 20	196
20 to 30	665
30 to 50	265

Table 4: Table to show emoji distribution of data with respect to age

Other examples of non dictionary words such as slangs or abbreviations not really present in Bengali, especially in written text and thus emojis were the only thing that could be analysed for the process.

That was followed by a process of cleaning the corpora as a whole to remove any unwanted characters that might hamper the analysis. This was done using regex and resulted in the preparation of a clean corpora for any further analysis.

This was followed by counting the number of words per sentence and plotting the frequency distribution of the same using a histogram. In addition to that, the distribution was also tabulated in buckets of 10 units and plotted as well to obtain a better insight into the distribution and frequency.

The results obtained across each of the categories are shown as follows.

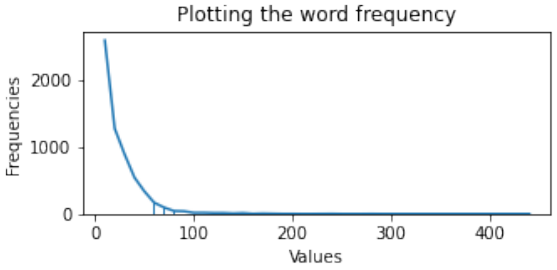

Figure 1: Gender: Male Word Frequency

This was followed by a process of word tokenization to identify the 30 most frequently

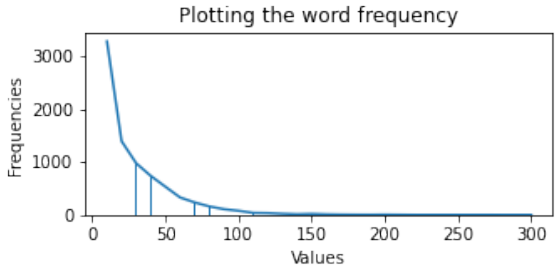

Figure 2: Gender: Female Word Frequency

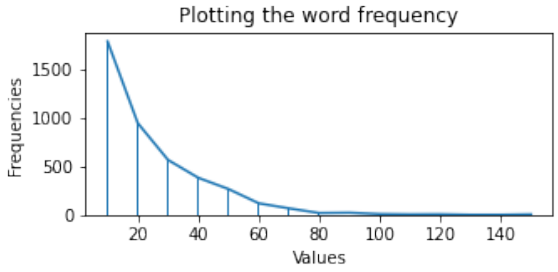

Figure 3: Age: 10 to 20 Word Frequency

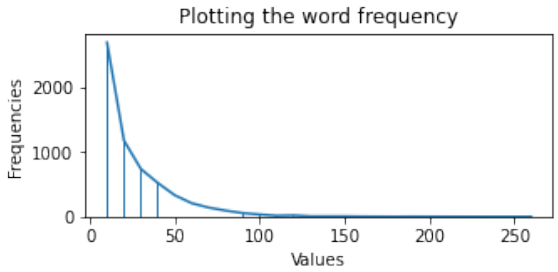

Figure 4: Age: 20 to 30 Word Frequency

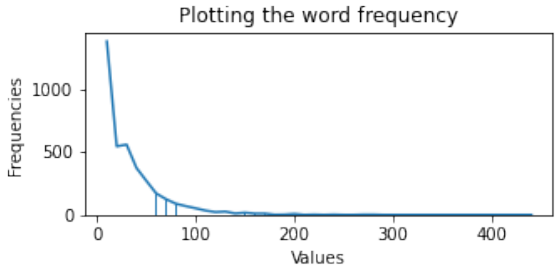

Figure 5: Age: 30 to 50 Word Frequency

present words. In addition to that, the corpora was further cleaned to remove any stop words and then the most frequently present words were plotted again. This will further be used in the process of building the word cloud later in the text analysis process.

Once the most frequent words were identified, the words were also give a POS tag and the most commonly occurring POS tags were also labeled - including and excluding stop words for the process of calculation and analysis.

The distribution of the 30 most commonly occurring POS tags was also plotted and the plots can be seen in the below images. *The reader might need to zoom in to be able to read and identify the labels of the graph better*

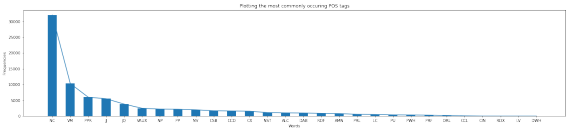

Figure 6: Age: 10 to 20 POS Frequency

Figure 7: Age: 20 to 30 POS Frequency

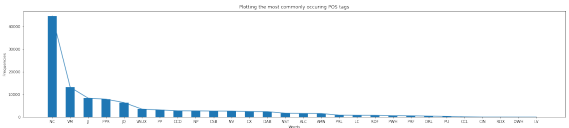

Figure 8: Age: 30 to 50 POS Frequency

Finally the words were stemmed, and a Word Cloud was created based on the given data to get a visual representation of the topics and important words that are commonly used in the text and if they vary across different age and gender categories.

After that, using a self designed algorithm which pays special attention to nouns and adjectives and then verbs, word clouds obtained. The picture of the same have been attached below. *(They use transliteration of Bengali text to aid the reader's understanding)*

In addition to all of this, during the process of carrying out the data analysis, the number of

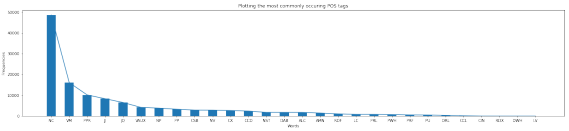

Figure 9: Gender: Male POS Frequency

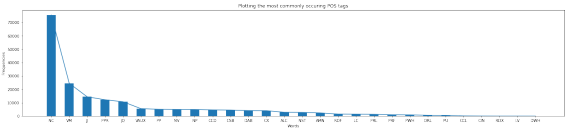

Figure 10: Gender: Female POS Frequency

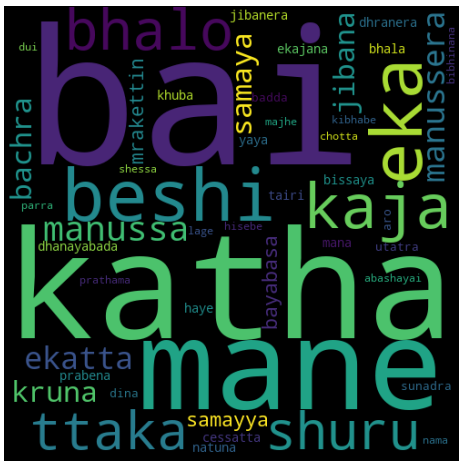

Figure 11: Gender: Male Word Cloud

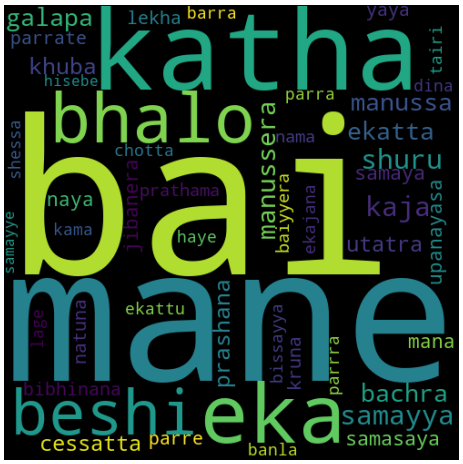

Figure 12: Gender: Female Word Cloud

Figure 13: Age: 10 to 20 Word Cloud

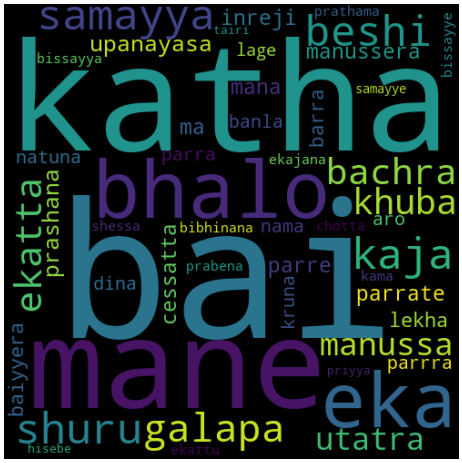

Figure 14: Age: 20 to 30 Word Cloud

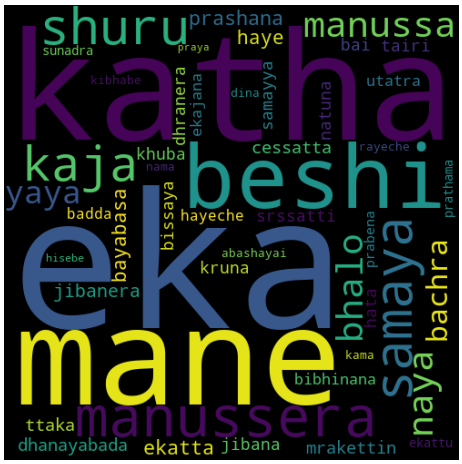

Figure 15: Age: 30 to 50 Word Cloud

words - including and excluding stop words and the number of sentences were also calculated which might help in coming up with better analysis of the data and the same has been tabulated as below.

Category	Words	Stopwords	Sentences
Male	123729	81481	6108
Female	190867	124534	7998

Table 5: Table to show word and sentence distribution of data with respect to gender

Category	Words	Stopwords	Sentences
10 to 20	78526	52456	4215
20 to 30	123775	81066	6090
30 to 50	112296	72495	3801

Table 6: Table to show word and sentence distribution of data with respect to age

The code used for the entire data analysis part of the project can be found in the linked Google Colaboratory Notebook. *(click on the name)*

3.3 Interpretation

As we can observe, the entire data analysis portion can be broken down into 4 major components - Emoji Count, Number of Words per sentence, Frequency of the different POS Tags and finally Word Cloud.

We will look into each of these factors individually in the upcoming section and try to obtain a better insight into the results that were observed in during the analysis.

3.3.1 Emoji Count

As we can see, normally it would seem that the 20 to 30 age category has a domination in the number of emojis and so do female users. However, we need to consider the fact that the corpora have varying number of words and thus a better analysis might be obtained by finding the ratio between the number of emojis and words.

In other words, what we have obtained here for a better understanding is the emoji to word percentage and the same has been tabulated in the given table.

From the above table, we can see how the in the case of gender both male and female users have similar percentages of emoji use, thereby showing how this criteria won't really be effective in analysing the gender category.

Category	Percentage
Male	3.66
Female	3.52

Table 7: Table to show emoji to word percentage with respect to gender

Category	Percentage
10 to 20	2.49
20 to 30	5.37
30 to 50	2.35

Table 8: Table to show emoji to word percentage with respect to age

On the other hand for the age categories, we can see how people in the 10 to 20 category have similar emoji use patterns as people in the 30 to 50 age category and the only people who have a significantly higher emoji use percentage are people belonging to the 20 to 30 category.

We can understand how people in their early twenties are thereby most likely to use emojis while writing something in Bengali. We need to however note that this is with respect to a informative, blog style writing website and the results might vary if things such as Twitter data was analysed instead.

In addition to that, the equitable distribution of emoji use among male and female users can also be justified by the reasoning of distribution of the users across different age categories, and more analysis can be done in order to find corelation between the same.

3.3.2 Number of Words per Sentence

When we look at the gender categories, we can see that the curve for male users sees a steeper decline compared to the curve for female users, which signifies that female users are likely to use more number of words per sentence as compared to their male counterparts.

We can also see that female users have quite a bit of cases with 100+ words per sentence, a phenomenon that is hardly observed in the case of the male users.

The reasoning behind the same might be the likelihood of female users using longer sentences to be more expressive with their speech

which as compared to male users who might want to make their text more concise and up to the mark instead of going on elaborating on their matter.

Looking at the data across age categories, we can see that there is a steeper decline in the curve for the 20 to 30 category as compared to the 10 to 20 category. However the 20 to 30 category has a pretty similar decline range as the 30 to 50 category, the only difference being how the distribution of 20 word and 30 word sentences are pretty similar for the latter while there is a pretty noticeable decline in the same for the former.

We can also notice how the maximum word count per sentence can go upto beyond 400 for the 30 to 50 category, while it reaches only upto 250 for the 20 to 30 category and only upto 150 for the 10 to 20 category.

Thus, we can see how the 10 to 20 age category is likely to use descriptive sentences but only upto a certain extent and their sentences hardly ever cross the 150 mark, which a user belonging to the 30 to 50 age category is more likely to do.

A similar reasoning as the one used for gender might be used in this case as well.

In addition to that, we can sort of hypothesize that as people grow older, they are more likely to write shorter sentences in general and thus their speech patterns progress more towards that of a male individual. However, there needs to be more research on this topic in order to come up with better and more conclusive results on the same.

3.3.3 POS Frequency

With respect to the gender categories it was seen that although nouns were found to reign superior, it was observed that female users are much more likely to use adjectives compared to prepositions while the observations were reversed for male users. Verb got the second position across the two different categories.

The reasoning as to why female users might be more inclined on using adjectives might be because they overall try to present data in a more expressive manner, while male users might try to resort to simple, concise presentation of data.

In the case of age categories, some interesting observations were found. Although nouns dominated the readings, it was seen that for the 10 to 20 age category they were followed by verbs, prepositions and adjectives; while for the 20 to 30 category they were followed by verbs, adjectives and prepositions. However, the most interesting observation comes from the 30 to 50 category, where nouns are followed by adjectives and then verbs and prepositions come much later in the ranking.

It is pretty hard to find a concrete reasoning so as to why it is the case and finding a proper reasoning without further, more directed analysis would not be accurate. However, the same reasoning of wanting to be expressive versus wanting to be concise might be used over here as well.

We can also notice now it might be harder to differentiate between the 20 to 30 versus 30 to 50 categories, while it might be much easier to differentiate between the two extremes in the domain of age categories.

3.3.4 Word Cloud

As we can see from the word clouds, the most common words for both male and female are largely similar and don't really have a significant variation in their distribution.

A similar pattern is also observed in the case of the distribution across age categories as well. We can observe how the top five words are more or less the same across all of these categories.

However, based on reading research papers written for the English Corpora and also because of the distribution of top writers in Quora as discussed in the Discussion section, we might be able to see a greater bias if we collect greater amounts of data and maybe even across writers who write over a wider range of topics.

I feel that might be able to show a better distribution of topics and the data in this case was insufficient to come up with a more concrete and better explainable conclusion in the case of the range of topics covered as it directly contradicts the distribution of writers across the topics.

4 Discussion

This following section initially states the observation made during data collection, which can give the reader a better insight into the process that went into collecting the data and the variation observed during that process.

A primary observation made was the distribution of the user base in Quora where at least in Bengali Quora, most users were found to be Male

and in the age category of 20 to 50, which made it difficult to find users in the other category. That was one of the factors which made the process of data collection even more tedious and time consuming.

As we can see there was also a limited number of writers in the age category of 10 to 20 as well, and most of them were in the upper realm of the age category, i.e., 15 and up.

What more was the fact that there were hardly any writers who fit the age category of 50 and above as a result of which that age category could not be included in the data for analysis. This can be explained by the fact that there is a reduced use in technology among that age group and the complexity of being able to type in another language apart from English while using technology as well.

Coming on to topics, it was seen that the top writers in most topics were dominated by males, and only a few topics had an equal distribution of female top writers - the topics being Writing and Bengali Literature.

All other categories, ranging from India, Books, Science to Technology, Photography as well as Human Life, Life and Living and Self Help were dominated by male writers - in some cases there was a complete absence of female writers in those categories as well.

A lot of these observations made can be explained when we look at the distribution of users of the internet according to age and gender where we observe that the highest user base is from the Male 20-50 demographic.

Coming onto the data analysis bit, most of the discussion has already been mentioned in the interpretation section where the results were analysed and a reasoning was formulated behind each of the mentioned observations that were obtained.

As final comments, we can see how Bengali is mostly noticed to be a pretty rigid language with respect to things such as the entry of slangs and non dictionary words and overall the way in which the language is used is mostly uniform across the different age and gender categories, apart from differences in the use of POS tags and thus the variety of words used.

Thus might be owing to the fact that most of the formal education in Bengali happens in a pretty uniform manner thereby resulting in it being carried over and maintained over the ages. In addition to that, another factor might be the fact that the only form of change that Bengali has mostly undergone in the recent past is the increased use of code mixing, something that is not observed as much in written text and thus it is hard to find differences across the patterns of language use in the different age categories.

Coming to the gender categories, as we have observed, there is a need of analysis of a larger corpora to account for the discussion on topics versus the results of the word cloud obtained, before we are able to reach more conclusive results.

5 Conclusion

As a final conclusion we can see how it is pretty hard to find differences in the use of language among Bengali users of different genders and ages albeit a few noticeable ones as enumerated here.

In the case of emojis, not much of a strong conclusion could be made apart from the fact that people belonging to the 20 to 30 age group are most likely to use emojis in their text. No such conclusions could be made in the gender category.

In the case of number of words per sentence, we can see how female users and younger users are more likely to use longer and more elaborated sentences as compared to the smaller and more precise sentences used by males as well as the older demographic.

Coming onto the factor of POS, we can again notice how females are more likely to use expressive text and thus use more adjectives. At the same time, the results obtained for the age categories were pretty interesting, showing how older people are also more likely to be more expressive compared to the younger users.

Finally, hardly any differences could be found in the top few words of the word cloud and thus it is hard to use that as a factor of difference. However, owing to the fact that there was a noticeable variation in the distribution of users across different topics, a larger corpora might be able to point out differences in the same.

6 Further Work

One of the most tedious processes for the project was data collection. It is hard to find anno-

tated corpora for Indian Languages, especially languages like Bengali where it is harder to obtain data.

Thus one of the main works that can be done in the future as an expansion of this project is to create a well annotated and labeled corpora for the purposes of authorship attribution and analysis, similar to the ones present in English.

In addition to that, had there been more time, it would have been possible to obtain more data from other places such as Twitter (which is arguably harder and more time consuming to annotate) and Blogs and use them for the purpose of this project as well.

Apart from that, it would also have been possible dig deeper in Quora and to collect data from more users, which will show more variety and help come up with better and more concrete results and would also remove the possibility of any bias that might creep into the data. Including data from a greater source of people would also help make better generalizations about the language group by the different social groups mentioned in this projects.

That would also enable more deep digging and being able to find data on the basis of more age categories and enable the expansion in that regard.

In addition to that, if more data is found and worked upon, and more linguistic variables are studied (maybe for a finer level of analysis), with enough data points, a Depp Neural Network Model can be created eventually which might be able to automatically classify any given Bengali data on the basis of age and gender.

The aforementioned model can be extremely helpful in the field of Forensic Linguistics, Bot Detection and can also help detect cases of Cyber-bullying, there by enabling better monitoring on the Internet.

References

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

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

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

Dipongkor A. K., Islam M. S., Kayesh H., Hossain M. S., Anwar A., Rahman K. A., Razzak I. 2021. DAAB: Deep Authorship Attribution in Bengali

Haque S., Ratul M. A. S., Khan M. Y. A. 2020. Open Source Autonomouns Bengali Corpus