BanglaMusicStylo: A Stylometric Dataset of Bangla Music Lyrics

Rafayet Hossain[1] Ahmed Al Marouf[2]
Department of Computer Science and Engineering
Human Computer Interaction Research Lab (HCI RL)
Daffodil International University (DIU)
Email: {rafayet3994[1], marouf.cse[2]}@diu.edu.bd

***Abstract*—With the rapid growth of Bangla music industry huge volume of Bangla songs are produced every day. Immense number of producers, lyricists, singers and artists are involved in production of songs from different genres. Among many genres of Bangla music; classical, folk, baul, modern music, Rabindra Sangeet, Nazrul Geeti, film music, rock music and fusion music has gained the highest popularity. Lyricists try to express their feelings and views towards any situation or subject through their writings. Therefore, each lyricist have their own dictionary of thoughts to put on music lyrics. In this paper, we have presented "BanglaMusicStylo", the very first stylometric dataset of Bangla music lyrics. We have collected 2824 Bangla song lyrics of 211 lyricists in a digital form. All the lyrics are stored in text format for further use. This dataset could be used for stylometric analysis such as authorship attribution, linguistic forensics, gender identification from textual data, Bangla music genre classification, vandalism detection, emotion classification etc. Identifying the significant research opportunities in this area, we have formalized this dataset which could be used for stylometric analysis.**

Keywords—Bangla Music Lyrics; Stylometric Analysis; Authorship Attribution; Bangla Stylometric Dataset.

I. INTRODUCTION

With the rapid growth of Bangla music industry, enormous amount of lyrics has been written by lyricists to produce music. Availability of music production tools and music sharing social networks such as YouTube, Vimeo, Amazon Prime Videos etc. are the main reason of great number of Bangla music production. Many peoples from different stages of society are involved in this industry. Lyricists, producers, singers and artists are involved with this entertainment industry. Apart from rhythm, tune, fusion, singer or genre; lyrics is the most vital element of songs, as it has direct impact on the listeners' choice and mood.

Bangla music genres could be divided into many sections, because of its highly diversified music categories. Genres could be categorized into three part: religious music (Hamd, Naat, Ghazal, Qawwali etc.), ethnic music (Baul, Bhatiali, Bhawaiya, Jari Gan, Sari Gan etc.) and traditional music (Rabindra Sangeet, Nazrul Geeti, Lalon, Hason Raja etc.). As people from different sector and different occupation are involved in Bangla music, this kind of diversity has adopted. For instance, the 'Bhatiali' is a genre of ethnic or folk music which is commonly sung by the boatmen in Bangladesh. The word 'Bhatiali' comes from 'Bhata' which means downstream. Choice of listening music depends on the lyrics of the songs. On the other hand, lyrics are form of text, which could be easily used in digital platforms for analysis. Text analytic tools could be proven as very effective tools to analyze the lyrics to find out the stylometric features of songs, for further applications.

Stylometry is the study of language style, especially on written form of language such as story, poems, and music lyrics. In the literature, researchers' have investigated the stylometric features or characteristics to find out the authorship attribution of Twitter data [1], analysis on scientific articles [2], to find the gender and age of bloggers [3], linguistic forensics [4] etc. It is evident that there could be many possible methodology to attribute authorship of documents, emails or story [5-8]. Therefore, to train a system properly to be used for other applications, a firm dataset. In Bangla computing or Bangla language processing the key challenge is lack of ground-truth dataset to train and test proposed methods.

In this paper, to the best of our knowledge, we presented the very first Bangla Stylometric dataset having 2824 Bangla lyrics of 211 lyricists. For our use in this paper, we would refer author or lyricist having same meaning. The rest of the paper is divided into six sections. Literature review is illustrated in section II, the attributes of the dataset are described in section III. Data collection procedure, statistical analysis of the dataset and applications of the dataset are demonstrated in section IV, section V and section VI, respectively. Finally, conclusion statements are in section VII.

II. LITERATURE REVIEW

This sections illustrates on the related works performed by the researchers in the Bangla computing area and associated area. This research area has got attention of researchers' from Bangla speaking countries specially Bangladesh and West Bengal of India. Million Song Dataset (MSD) [17] is an English song dataset containing one million song, which is a cluster of complementary datasets having cover songs, lyrics, song-level tags, user data, genre labels etc. musiXmatch[1] is the official lyrics dataset of MSD having 237,662 tracks lyrics. Similar datasets could be found having English songs, but not available for Bangla songs. Song lyrics could be used for analyzing many research topics such as emotion classification [11-13, 16], mood classification [14], semantic analysis [15] etc. A. Jamdar et al. [13] proposed a lyrical and audio features

[1]https://labrosa.ee.columbia.edu/millionsong/musixmatch

978-1-5386-8207-4/18/$31.00 ©2018 IEEE

based method to detect the emotion of a song. He applied ANEW [19] and WordNet [18] knowledge to associate the linguistic features extracted from lyrics. Weighted and stepwise threshold reduction on KNN algorithm is applied for the classification task. R. Malheiro et al. [16] created a song dataset containing 180 song lyrics according to Russell's emotion model. He extracted features complemented by stylistic, structural and semantic features to identify the arousal and valence categories of each song. In [16] regression analysis and different criteria-wise classification is also applied.

X. Hu et al. [14] proposed a text mining method to classify the music mood. He proposed a ground-truth database of English songs having total 21,000 songs. Among these songs only 8784 English songs have lyrics and the proposed method applies WordNet-Affect [20], a linguistic resource to filter the affective meanings of the tags. The method uses Bag-of-Words (BoW), Part-of-Speech (POS) and function words as features and SVM as classifier. B. Logan et al. [15] performed a semantic analysis on song lyrics on the publicly available *uspop2002* dataset [21]. He applied Probabilistic Latent Semantic Analysis (PLSA) on the dataset. Moreover, this paper also focuses on the artist similarity, acoustic similarity and topic modeling. Natural language processing (NLP) could be applied on lyrics [22] collected from Lyrics.com[2] and Lyrics4u.com[3]. Mahedero et al. [22] proposed a language identification, structure extraction and thematic categorization methods.

In this paper, we have proposed the very first Bangla Stylometric dataset '*BanglaMusicStylo*' which could be used for many further application including authorship attribution, linguistic forensics etc. This paper could be considered as the starting journey towards the exploration of research possibilities in Bangla stylometric analysis.

III. Dataset Attributes

In *BanglaMusicStylo* dataset, the lyrics of 2824 Bangla songs are stored. In this collection, we have tried to cover the most popular genres of Bangla music. Some attributes of the dataset could be listed as following.

- Different authors song lyrics are kept in separate folders.
- Different songs of same author are kept in the same folder.
- Different song lyrics are stored in '*Siyam Rupali*' Bangla font in Microsoft .docx file format. Using simple file reader methods in Java or any object oriented programming language, it is possible to read the separate lines of the files for further text processing.

[2]https://www.lyrics.com/
[3]http://lyrics4u.com/

Fig. 1. Author folders having songs inside.

The Fig. 1 illustrates some of the author folders and snippet of the dataset. Fig. 2 shows the song files named as "songID_songTitle.docx" format. Each file contains the lyrics of the songs. The songs are written by the national poet of Bangladesh, Kazi Nazrul Islam.

Fig. 2. List of songs of Kazi Nazrul Islam.

Each file contains lyrics in text format, written in *Siyam Rupali* Bangla font. The lyrics also contains the necessary notations used to support the singers to sign the song properly, such as number of repetition, specification of song sections.

In this dataset, we have collected the songs from different genres of Bangladeshi music. We tried to collect song lyrics from each category of religious music, ethnic music and traditional music. Fig. 3 shows an example song lyrics of Rabindranath Tagore, the writer of national anthem of Bangladesh.

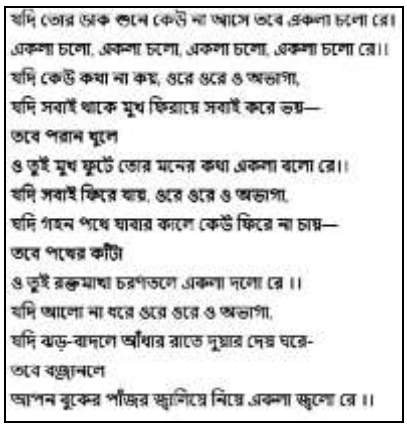
যদি তোর ডাক শুনে কেউ না আসে তবে একলা চলো রে।
একলা চলো, একলা চলো, একলা চলো, একলা চলো রে।।
যদি কেউ কথা না কয়, ওরে ওরে ও অভাগা,
যদি সবাই থাকে মুখ ফিরায়ে সবাই করে ভয়—
তবে পরান খুলে
ও তুই মুখ ফুটে তোর মনের কথা একলা বলো রে।।
যদি সবাই ফিরে যায়, ওরে ওরে ও অভাগা,
যদি গহন পথে যাবার কালে কেউ ফিরে না চায়—
তবে পথের কাঁটা
ও তুই রক্তমাখা চরণতলে একলা দলো রে ।।
যদি আলো না ধরে ওরে ওরে ও অভাগা,
যদি ঝড়-বাদলে আঁধার রাতে দুয়ার দেয় ঘরে-
তবে বজ্রানলে
আপন বুকের পাঁজর জ্বালিয়ে নিয়ে একলা জ্বলো রে ।।

Fig. 3. Snippet of the song "যদি তোর ডাক শুনে কেউ না আসে" authored by Rabinranath Tagore.

Most of the songs of the proposed dataset could be classified in the following genres of Table I.

TABLE I. BANGLADESHI MUSIC GENRES

Genres	Description
Classical	Classical music is based on modes called rags. Based on various versions of Hindustani classical forms Bangla classical music are adopted.
Folk	This genre distinguished by simple musical instruments and words. It has evolved from the traditional cultures.
Baul	It most commonly known category of Bangladeshi folk songs. It incorporates simple words expressing song with deeper meanings. These songs are performed with very little musical instruments such as '*aktara (one string instrument) or dotara (two string instrument)*' and supports the main vocal.
Adhunik or Modern Music	Contemporary songs are generalized as *adhunik* or modern music. Nowadays, newbie singers are mainly focusing in producing this genre songs.
Rabindra Sangeet	It also known as "Tagore Song". Written and composed by Rabindranath Tagore, the Nobel Prize winning Bengali writer, who wrote the national anthem of Bangladesh as well as India. People like the overall tone, rhythm and lyrics of these songs for more than decades.
Nazrul Geeti	Nazrul geeti songs are written and composed by Kazi Nazrul Islam, a famous Bengali poet and national poet of Bangladesh. He is specially known because of his revolutionary poems which are converted to songs.
Film music	The film industries of Bangladesh supported music by according reverences to classical music while utilizing the western orchestration to support melodies. Film music are placed in the films to decrease monotonous story line into interesting one. These genres consists romantic, sad, happy and anger emotions songs.
Rock music	Bangladeshi rock music was introduced in 1972 by a singer, song writer, composer Nasir Ahmed Apu. Influenced by western music, Bangladeshi young artists got involved in this trend and produced some of the most popular songs.
Fusion music	Traditional music with western instrument to revitalize and re-popularize Bengali music. This genre is recently becoming popular to the young listeners.

IV. DATA COLLECTION PROCEDURE

The '*BanglaMusicStylo*' dataset contains 2824 Bangla song lyrics. For collecting these Bangla song lyrics, we have used meta-searching techniques. We have used keyword based searching in the World Wide Web (WWW). We have picked up numerous keywords to search the lyrics from sites, blogs or in audio or video sharing sites. The keywords are chosen from different categories such as lyricists name, song titles, song genres, emotion words. But our main focus was to collect as much as song lyrics possible for each lyricist. In this dataset we have collected 211 lyricists song. Among them more than 10 songs are collected of 38 lyricists. The keywords for searching criteria are shown in Table II. We have used both Bangla and English keywords to collect the data.

TABLE II. KEYWORDS USED FOR META SEARCHING

Keyword Category	Example Keywords
Name of the Lyricist	Rabindranath Tagore, Kazi Nazrul Islam, Gazi Mazharul Anowar, Lalon, Gauri Prasanna Majumder, Hason Raja etc.
Title of the songs	মায়াবন বিহারিনী, আমার পরান যাহা চায়, নেশা লাগিলো রে, সবাই তো সুখি হতে চায়, আমার সোনার বাংলা, তারায় তারায় etc.
Genre class	Bangla Rock song, Bangla Band song, Rabindra Sangeet, Nazrul Geeti, Bangla Folk song list, Hason Raja songs etc.
Emotional words	Bangla sad songs, Bangla celebration songs, Bangla happy songs etc.

Apart from the keyword based searches, we have also collected lyrics from the album covers and comment section of YouTube. Many listener try to comment the song lyrics under the YouTube videos if they like the song. We have also gathered some of the lyrics from there, which are mostly modern music.

V. STATISTICAL ANALYSIS OF DATASET

In this section, we have illustrated the statistical perspective of the '*BanglaMusicStylo*'. Table III shows the properties of the dataset and Table IV demonstrates some of the insights of number of songs per author. The average songs per author and more than thousand words per author would be sufficient to train-test a machine learning system applying text mining algorithms.

TABLE III. PROPERTIES OF THE DATASET

Properties	Values
Total Number of Songs	2284
Total Number of Words	224,342
Avg. Songs per Author	13.38388626
Avg. Words per Author	1063.232227

In Table IV, the snippet of author-wise number of songs and words are listed. The highest number of songs (856) and second highest (620) are collected of Rabindranath Tagore and

Kazi Nazrul Islam, the two most influential lyricists in Bangla music. The dataset contains 20 plus songs of 15 lyricists and 10 plus songs of more than a hundred lyricists. More than a thousand words of lyrics are collected for 100 lyricists.

TABLE IV. SNIPPET OF THE AUTHOR-WISE DATA STATISTICS

Author Name	No. of Songs	No. of Words
Rabindranath Tagore	856	52784
Kazi Nazrul Islam	620	43246
Gazi Mazharul Anwar	82	7282
Pulak Bandyopadhyay	66	5955
Gauri Prasanna Majumder	62	5080
Lalon Shah	38	3108
Latiful Islam Shibli	30	3052
Shibdas Bandyopadhyay	24	1773
Kabir Bakul	22	2483
Mohammad Rafiquzzaman	22	1559

VI. APPLICATIONS OF DATASET

In this section, we described the applications of the proposed dataset. Using this dataset could be applied to the following challenging tasks, but are not limited to.

A. Authorship Attribution

Authorship attribution is the most common approach to identify the author of the written document or music [1]. This algorithm has been applied in many contexts such as document, story, music lyrics or even social media text data like Twitter data [1]. The proposed dataset could be used for identifying the authors of Bangla Music lyrics.

B. Linguistic Forensics

Linguistic forensics is an ancient method to recognize the gender, age or other characteristics of author [4]. Finding sufficient features from the text applying text mining tools and classify the features into gender or age of the author. With this dataset we have the derived gender of the lyricists, therefore, this dataset could be used for the same purpose.

C. Bangla Music Genre Classification

Ashfaqur et al. [9] proposed a Bangla music genre classification method based on learning and prediction approach for classifying four genre of songs namely, Rabindra Sangeet, Folk Song, Modern song and pop music. Our dataset could be used for similar purpose as we have collected song lyrics of nine different genre mentioned in Table I.

D. Vandalism Detection

Vandalism detection considered as one-class classification problem applied on textual data [10]. Character level, word level and sentence level features could be used as content features for the classification task. Our proposed dataset could be used for vandalism detection as it contains the lyrics as text.

E. Emotion-based Classification

Emotion based text classification [11] is a challenging task to classify text content based on the understandings of emotion cues from it. Finding emotion from US song lyrics is presented in [12] based on linguistic markers of psychological traits and emotions over time. Similar emotion based classification could be adopted for Bangla song lyrics using our proposed dataset.

VII. CONCLUSION

This paper proposed a stylometric dataset of Bangla song lyrics for analysis of stylometric features. To the best of our knowledge, this is the very first Bangla song lyrics dataset which could be used in many stylometric analysis. As Bangla computing is extending its branches in many dimensions, stylometric analysis of Bangla lyricists is a considerable task to be performed. This paper investigates the possibilities of future research in Bangla stylometry.

REFERENCES

[1] M. Bhargava, P. Mehndiratta and K. Asawa, "Stylometric analysis for authorship attribution on twitter.", In Big Data Analytics, vol. 8302 of Lecture Notes in Computer Science, pp. 37–47. Springer International Publishing, 2013.

[2] S. Bergsma, M. Post and D. Yarowsky, "Stylometric analysis of scientific articles.", North Americal Chapter of Association of Computational Linguistics (NAACL), 2012.

[3] S. Goswami, S. Sarkar and M. Rustagi, "Stylometric analysis of bloggers' age and gender", In Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM), San Jose, USA, May 17 - 20, 2009.

[4] M. T. Turell, "The use of textual, grammatical and sociolinguistic evidence in forensic text comparison", International Journal of Speech Language and the Law, vol. 17, issue. 2, 2010.

[5] J. Diederich, J. Kindermann, E. Leopold and G. Paass, "Authorship attribution with Support Vector Machines", Applied Intelligence, vol. 19, pp. 109–123, 2000.

[6] H. Craig, "Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them?", Literary and Linguistic Computing, vol. 14, issue. 1, pp. 103–113, 1999.

[7] A. Gray, P. Sallis and S. MacDonell, "Software forensics: Extending authorship analysis techniques to computer programs.", 3rd biannual conference of the International Association of Forensic Linguists (IAFL), 1997.

[8] D. Lowe and R. Matthews,"Shakespeare vs. Fletcher: A stylometric analysis by radial basis functions.", Computers and the Humanities, vol. 29, pp. 449–461, 1995.

[9] A. Rahman, "Bangla Music Genre Classification", Journal of Multidisciplinary Computational Intelligence Techniques: Application in Business, Engineering and Medicine, pp. 15, 2012.

[10] S. Heindorf, M. Potthast, B. Stein and G. Engels, "Vandalism Detection in Wikidata", Conference on Information and Knowledge Management (CIKM), Indianapolis, USA, October 24-28, 2016.

[11] T. Danisman and A. Alpkocak, "Feeler: Emotion classification of text using vector space model.", In AISB 2008 Convention, Communication, Interaction and Social Intelligence, vol. 2, pp. 53–59, Aberdeen, Scotland, 2008.

[12] C. N. DeWall, R. S. Pond, W. K. Campbell and J. M. Twenge, "Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular U.S. song lyrics.", Psychology of Aesthetics, Creativity, and the Arts, vol. 5, pp. 200-207, 2011.

[13] A. Jamdar, J. Abraham, K. Khanna and R. Dubey, "Emotion Analysis of Songs based on lyrical and audio features", International Journal of Artificial Intelligence & Applications (IJAIA), vol. 6, issue. 3, May 2015.

[14] X. Hu, J. S. Downie and A. F. Ehmann, “Lyric text mining in music mood classification”, 10th International Society for Music Retrival Conference (ISMIR), 2009.

[15] B. Logan, A. Kositsky and P. Moreno, “Semantic Analysis of Song Lyrics”, IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, 27-30 June, 2004.

[16] R. Malheiro, R. panda, P. Gomes and R. P. Paiva, “Emotionally-relevant features for classification and regression of music lyrics”, IEEE Transactions of Journal Affective Computing, 2016.

[17] T. B. Mahieux, D. P. W. Ellis, B. Whitman and P. Lamere, “The million song dataset”, International Society for Music Information Retrieval, 2011.

[18] G. A. Miller, “Wordnet: A lexical database for English.”, Community of ACM, vol. 38, issue. 11, 1995.

[19] M. M. Bradley and P. J. Lang, “Affective norms for English words (ANEW): Instruction manual and affective ratings”, (Tech. Rep. No. C-1). Gainesville, FL: University of Florida, The Center for Research in Psychophysiology, 1999.

[20] C. Strapparava and A. Valitutti, “WordNet-Affect: an Affective Extension of WordNet,” Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp. 1083-1086, 2004.

[21] A. Berenzweig, B. Logan, D.P.W. Ellis and B. Whitman, “A large-scale evaluation of acoustic and subjective music similarity measures.”, In Proceedings International Conference on Music Information Retrieval (ISMIR), 2003.

[22] J. P. G. Mahedero, “Natural language processing of lyrics.”, In Proceedings of the 13th annual ACM International conference on Multimedia, New York, USA, 2005.