A Stylometric Dataset for Bengali Poems

Mirza Kamrul Bashar Shuhan
bKash Limited
Dhaka, Bangladesh
shuhan.mirza@gmail.com

Rupasree Dey
Oregon State University
Corvallis, Oregon, United States
rupasreedey5@gmail.com

Sourav Saha
Shahjalal University of Science and Technology
Sylhet, Bangladesh
souravsaha201@gmail.com

Md Shafa Ul Anjum
Progoti Systems Limited
Dhaka, Bangladesh
shaown229@gmail.com

Tarannum Shaila Zaman
SUNY Polytechnic Institute
Utica, New York, United States
zamant@sunypoly.edu

ABSTRACT

Poetry is a form of literature that conveys feelings using different styles, aesthetics, and rhythms. The Bengali language has an enriched collection of poems. Every poet has an individual style of expressing their thoughts and emotions. However, stylometric research in this branch of the Bengali language is still in its early stage of development. In this paper, we have presented a stylometric dataset, which has 6,070 poems of 137 poets stored in the textual format. To the best of our knowledge, this is the first stylometric dataset for Bengali poems which will add an extra dimension to the expanding research arena of the Bengali language. To explore the usability of this dataset, we developed poem genre classifiers using deep learning that can classify these poems. Performance analysis of some deep learning classifiers has been presented in addition to classification. The classifiers include GRU and CNN. Among these two, GRU showed better performance by 91.48 in terms of the F1-score. The dataset will be publicly available at https://github.com/shuhanmirza/Bengali-Poem-Dataset after publishing this article.

CCS CONCEPTS

• **Computing methodologies → Language resources**.

KEYWORDS

Natural Language Processing, Language Resources, Datasets, Stylometric Datasets, Bengali Poems, Corpora

ACM Reference Format:
Mirza Kamrul Bashar Shuhan, Rupasree Dey, Sourav Saha, Md Shafa Ul Anjum, and Tarannum Shaila Zaman. 2022. A Stylometric Dataset for Bengali Poems. In *2022 6th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2022), December 16–18, 2022, Bangkok, Thailand.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3582768.3582788

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
NLPIR 2022, December 16–18, 2022, Bangkok, Thailand
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9762-9/22/12…$15.00
https://doi.org/10.1145/3582768.3582788

1 INTRODUCTION

At this moment, Bengali is the sixth language in terms of the number of speakers, as over 268 million people all over the world speak this language[14]. As far as we are aware, there is no poetry dataset to perform the stylometric research in Bengali. In this paper, we aimed to develop one such poem dataset. This dataset can be treated as an enriched stylometric resource to perform different tasks like authorship attribution, genre classification, automatic poem generation, sentiment analysis, and many more. With the improvement of machine learning and deep learning models, narratives are now more interpretable to computers. The latent patterns of the narratives are analyzed with the help of the sophisticated architecture of the algorithms. Because poetry is one of the most difficult forms of narrative to grasp by a computer due to its complicated syntax, a dataset containing a vast variety of poems can aid in the process of exploring hidden patterns in poems using machine learning or deep learning models.

In recent times, although there has been a significant number of works in informal text-domain like customer reviews, social media posts, and conversations, the number of works dealing with formal texts is very limited. Formal texts include novels, poetry, plays, official documentation, etc. Therefore, one motivation to work on a stylometric dataset of Bengali poems came from here. To talk about the impacts of this research, meaning or sentiment extraction, automatic poem generation is one stylometric task that can be facilitated by using this dataset. The models built and trained on this dataset can be used to build an automatic poem generation system that can be treated as an excellent tool for children learning. Moreover, authorship attribution can also be done based on this poem corpus where the writing patterns of different authors can be analyzed. As this dataset contains poems by more than 130 poets of Bengali literature, understanding their writing patterns could be a milestone in the authorship attribution domain with long formal texts with complex sentence structures.

Our contributions can be conferred as follows.

- We prepared a novel dataset, containing 6,070 poems from 137 poets, that can be treated as a legitimate source to dive into the stylometric analysis of Bengali poems.
- Tagging classes to the poems is an additional part of this work. The classes correspond to the structure and inherent meaning of the poems. The total number of classes is 21.

Table 1: Properties of the Dataset

Properties	Value
Number of poets	137
Total number of poems	6,070
Total words	1,011,477
Total sentences	173,107

Table 2: Number of poems in various classes

Love/Erotic	1356
Reflective	1316
Humanism	723
Sonnet	468
Devotional	394
Allegory	392
Nature	347
Rhymes	323
Patriotic	249
Didactic	180

Comedy	127
Narrative Story	57
Elegy	55
Lyrical Ballad	31
Satire	21
Ode	10
Epic	9
Ballads	4
Dramatic	3
Epistle	3
Lyrical Drama	2

- To show a field where this dataset can be impactful, poem classification has been performed using the model trained on the poems in the dataset.

The following sections of the paper confer on the dataset preparation method, statistical analysis of the dataset, dataset annotation procedure, and applications of the dataset.

2 DATASET ATTRIBUTES AND STATISTICAL ANALYSIS

The stylometric dataset for Bengali poems is presented here from a statistical point of view. We have collected 6,070 poems in total belonging to 137 poets in this work. The primary statistical attributes of this collection are listed in table 1. Besides that, table 2 and table 3 show the count of poems per class and poems per poet to illustrate a better understanding of the dataset. From table 2, it is visible that the most prominent poem class is *Love* having 1356 poems. The closest class is *Reflective* with a 1316 poem count. These two classes constitute around 45% of the total dataset. 9 out of 21 classes have poem counts in between 100 and 1000. They constitute around 53% of all the poems. The least prominent classes are Epic, Ballads, Dramatic, Epistle, and Lyrical Drama where all of which have a poem count of less than 10.

To describe table 3, the highest number of poems belong to Rabindranath Tagore. The second-highest number of poems are written by Shamsur Rahman. Around 43% of all the poems are taken from these two poets.

Table 3: Number of poems by poets (who have more than 100 poems)

Rabindranath Tagore	1503
Shamsur Rahman	1075
Jibanananda Das	390
Kazi Nazrul Islam	383
Mahadev Saha	239
Nirendranath Chakravarty	187
Purnendu Pattrea	184
Sunil Gangopadhyay	163
Joy Goswami	158
Sukumar Ray	153
Sukanta Bhattacharya	135
Jasimuddin	129
Michael Madhusudan Dutt	129

3 DATASET PREPARATION AND ANNOTATION

3.1 Data Collection

The poems have been collected through web scraping of various poem blogs and web resources. These sources are [3],[5], [4], [9] and [2]. We intended to explore all possible websites and blogs where we could find both a legitimate source and a significant number of poems. However, the data collected from these sources were unstructured and needed cleansing.

3.2 Data Cleansing

The scraped data had many unnecessary HTML tags and characters that made the data noisy. The major pre-processing tasks included removing HTML tags, unnecessary spaces, newlines, and some unexpected punctuation marks. Furthermore, we removed all incomplete and duplicate entries and converted the poem contents into a UTF-8 format. The data was cleansed through both automation scripts and manual effort.

3.3 Organizing Dataset

Upon cleansing the data, we organized the data and built the stylometric dataset. Every poem entry in the dataset includes; (a) poem name, (b) poem content, (c) poet name, (d) poem class, (e) source, and (f) timestamp of scraping.

3.4 Class Tagging of Poems

A poem can express a variety of emotions and messages in diverse structures. It is quite impossible to put a poem under a single class. However, we can tag a class to a poem if that poem shows strong consistency with the properties of that class. In "Sahitya Sandarshan (A glossary of literary terms)" by [7], we have seen the writer has analyzed Bengali poetry formally. The poems can be divided into major two classes, (1) Subjective and (2) Objective. These classes have multiple sub-classes. However, according to the formal analysis of Bengali poems, a poem can never be fully Subjective or Objective. We have put definitions and properties of various poem classes in appendix A.

We have tagged the poems of our dataset following the work of [7]. This book is considered the most fundamental analysis of Bengali literature.

4 POEM CLASSIFICATION

Before feeding data into the deep learning models, the following steps were applied for all the models. To reduce the bias of the majority classes, oversampling by duplicating the samples of minority classes was performed. The other pipeline tasks include label encoding, dataset splitting, and tokenizing.

4.1 Model Description

4.1.1 Recurrent Neural Network(GRU unit). The Gated Recurrent Unit [6] has been used as one of the classification models in this paper. The values of the hyperparameters of this model are listed in table 4.

Table 4: hyperparameters values of GRU model

Hyperparameter	Value
Maximum Sequence Length	52
Embedding Layer Vocabulary Size	106728
Embedding Dimension	64
Dropout	0.2

The model was trained up to 30 epochs because the accuracy was remaining steady after 30 epochs. Each epoch contains 64 batches.

4.1.2 Convolutional Neural Network. Convolutional Neural Network(CNN) [10] has also been used as a classification model in this work. The values of the hyperparameters of this model are listed in table 5.

Table 5: hyperparameters values of the CNN model.

Hyperparameter	Value
Maximum Sequence Length	52
Embedding Layer Vocabulary Size	106728
Kernel Size	5
Number of filters	512

The model was trained up to 30 epochs because the accuracy was remaining steady after 30 epochs. Each epoch contains 64 batches.

Going through the appendix section B of this paper can give a better understanding of the model architectures of both GRU and CNN.

4.2 Performance Analysis of the Classifiers

The accuracy score used in this paper includes precision, recall, F1-score, and support. These accuracy scores for two classifiers have been presented in the table 6.

Table 7 shows the number of properly and misclassified examples for both classifiers.

Both of the classifiers fail to classify some examples of some particular classes. For GRU, the classes where the mistakes are visibly frequent are Allegory, Nature, Ode, and Ballad. For these four classes, the number of misclassified examples is 179, 74, 107, and 47 where the total number of poems for each class is 272. Using CNN, the classes where the mistakes are visibly frequent are Metaphoric, Nature, Ode, and Ballad. For these four classes, the number of misclassified examples is 150, 79, 110, and 32 where the total number of poems for each class is 272. For the GRU classifier, accuracy change over time is shown in the table 8.

For both of the classifiers, the classes that have fewer poem samples, have more misclassified examples. Undersampling the dataset might be a solution to reduce the number of misclassified examples for those classes.

Table 6: Accuracy of the classifiers

Classifier Name	Precision	Recall	F1-score	Support
GRU	91.48	91.48	91.48	0.914797
CNN	90.407	90.43	90.48	0.907

Table 7: Correctly and incorrectly classified examples

Classifier	Correctly Classified examples	Misclassified examples
GRU	5,218	494
CNN	5,119	593

Table 8: Test Accuracy over Time(seconds) for GRU and CNN

Time	Accuracy
200	86
600	91
800	91.48
1200	90.20

Time	Accuracy
200	87.55
600	90.02
800	90.407
1200	90.20

5 DISCUSSION

We have presented a novel stylometric dataset of Bengali poems in this paper. This dataset can be leveraged in many research fields such as Authorship Attribution, Sentiment Analysis, and, Linguistic Forensics. To explore the usability of the dataset, we analyzed the performance of some state-of-the-art classifiers operating on the dataset.

However, the dataset represents a fraction of the vast Bengali poetry. We see this work as an initial effort to build a significant corpus of Bengali poetry.

6 FUTURE WORK

We intend to expand the dataset by adding a vast amount of poems to it as well as incorporating more properties of poems like publishing year, publisher's name, book's name, and rhythm. To facilitate the participation of the public, we want to upload and maintain the dataset in public git repositories like Github. This will enable enthusiasts to contribute to the dataset with ease and transparency.

Additionally, we want to make use of the dataset in our future research works such as authorship attribution, automatic poem

generation, and poem classification. Transformer-based methods like BERT [8] can be considered in the follow-up works to classify poems.

7 RELATED WORKS

The English language has several publicly available poem corpus. Gutenberg Poetry Corpus [11] is one of the notable sources of English poems, which contains approximately three million lines of poetry, extracted from hundreds of books. The corpus is particularly fit for applications in creative computational poetic text generation. [12] owns another English poetry database, which contains over 160,000 poems, by around 1,250 poets.

Similar to the dataset preparation field, poem classification is another field where very few works have been done in the Bengali language. Among them, [13] used a Multiclass SVM classifier to classify Tagore's collection of poetry into four categories: devotional, love, nature, and nationalism. They conclude that for poetry classification, using word features is not sufficient because allusions often being used as a poetic device. Because of the lack of work in Bengali, we have explored classification works in other languages as well. Among them, [1] proposed an attention-based C-BiLSTM model to classify the text of poetry into different emotional states, like love, joy, hope, sadness, anger, etc.

Exploring the related works in Bengali, we found that all of them created a collection of poems for their research but did not publish them. So our objective in this paper was to publish a source that can be utilized in the future by those who are interested in various stylometric research works.

8 CONCLUSION

The primary objective of this paper was to present a novel stylometric dataset of Bengali poems. The dataset incorporates over 6000 poems from more than 130 poets. A comparison of several deep learning models to classify poems was shown in section 4, exploring the usability of the dataset. Among the two models experimented with, GRU performed better than CNN with a 91.48 F1-score. This dataset will facilitate stylometrically analytic tasks like authorship attribution and sentiment analysis in the future.

ACKNOWLEDGMENTS

We thank the bKash Limited for funding this research partially.

REFERENCES

[1] Shakeel Ahmad, Muhammad Zubair Asghar, Fahad Mazaed Alotaibi, and Sherafzal Khan. 2020. Classification of poetry text into the emotional states using deep learning technique. *IEEE Access* 8 (2020), 73865–73878.
[2] Bangla-kobita. [n. d.]. Bangla-kobita. https://bangla-kobita.com. [Online; accessed 10-April-2022].
[3] Banglakosh. [n. d.]. Banglakosh. https://http://www.banglakosh.com/. [Online; accessed 10-April-2022].
[4] Banglapoems. [n. d.]. Banglapoems. https://banglapoems.wordpress.com. [Online; accessed 10-April-2022].
[5] Banglarkobita. [n. d.]. Banglarkobita. https://banglarkobita.com. [Online; accessed 10-April-2022].
[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
[7] Shrishchandra Dash. 1960. *Sahitya Sandarshan (A glossary of literary terms).*
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[9] Kobitacocktail. [n. d.]. Kobitacocktail. https://www.kobitacocktail.com. [Online; accessed 10-April-2022].
[10] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[11] Allison Parrish. [n. d.]. A Gutenberg Poetry Corpus. https://github.com/aparrish/gutenberg-poetry-corpus. [Online; accessed 15-April-2022].
[12] Princeton-University-Library. [n. d.]. English Poetry Database. https://library.princeton.edu/resource/3772. [Online; accessed 2-September-2022].
[13] Geetanjali Rakshit, Anupam Ghosh, Pushpak Bhattacharyya, and Gholamreza Haffari. 2015. Automated analysis of bangla poetry for classification and poet identification. In *Proceedings of the 12th International Conference on Natural Language Processing*. 247–253.
[14] Statista. [n. d.]. Most Spoken Languages in the World. https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide. [Online; accessed 14-February-2022].

A CLASSES OF BENGALI POEM

Bengali poems are divided into two major classes; (1) Subjective and (2) Objective. These classes can further be divided into various sub-classes [7].

Major categories of Subjective Poetry are; (a) Devotional, (b) Nature, (c) Sonnet, (d) Lyrical Drama, (e) Dramatic Lyric, (f) Love or Erotic, (g) Patriotic, (h) Elegy, (i) Humanism, (j) Reflective, (k) Ode, (l) Lyrical Ballad, and (m) Rhyme.

Similarly, major sub-classes of Objective poetry are ; (a) Satire, (b) Epistle, (c) Allegory, (d) Didactic, (e) Narrative Story, (f) Ballads, (g) Comedy, and (h) Epic.

The brief definitions of the major classes are presented below.

- Love or Erotic: This category includes two types of poems, union, and separation between the two people who are in a relationship. Besides that, love for a particular person of attraction may also be expressed.
- Reflective: Usually, the states of poets are expressed through different entities like images or symbols and the poet makes a comparison between the imaginary world and the real world based on images or scenes.
- Devotional: The poems typically contain devotional expressions to God or any specific person.
- Patriotic: Love for the motherland is a major indicator to identify this category. Distress or suffering of the native people is also an indicator.
- Elegy: Poets convey their feelings of grief toward any living entities or even objects in these types of poems.
- Nature: The description of natural elements or love for them is shown here.
- Sonnet: This kind of poem belong to a specific structure where the total number of lines is 14 and the letters of each line are 14 as well.

B DESCRIPTION OF CLASSIFICATION MODELS

The following sub-sections illustrate on the model architecture and hyperparameter settings of the classification models.

B.1 Recurrent Neural Network(GRU unit)

Gated Recurrent Unit [6] shows prominent performance in text-data categorization. The layers of the model as well as the hyperparameters used in our experiment are described below.

- Embedding layer: When the text data is converted into integers or numerical representation, it can be applied as input to the deep learning model. In the Embedding layer, each integer is converted into vector form. The parameters of this layer are 106728 as vocabulary size, 64 as embedding dimension, and 52 as maximum length.
- GRU Layer: This layer has 64 units and the dropout value is set to 0.2.
- Dense Layer: This is a fully connected dense layer having 64 units and the activation function used is 'Relu'.
- Flatten Layer: Multi-dimensional output created in the previous layer is converted into a single-dimensional form in this layer.
- Output Dense Layer: The final output layer is a fully-connected dense layer that has some units that are equal to the total number of classes in our dataset, which is 21. Softmax is used as the activation function here and it computes the probability of the classes.

B.2 Convolutional Neural Network

Convolutional Neural Network(CNN) [10] which has been used mostly for image classification purposes, is also being used in text classification tasks.

The layers of the model as well as the hyperparameters used in our experiment are described below.

- Input Layer: This layer has a maximum sequence length value of 52.
- Embedding layer: When the text data is converted into integers or numerical representation, it can be applied as input to the deep learning model. In the Embedding layer, each integer is converted into vector form. The parameters of this layer are 106728 as vocabulary size, 64 as embedding dimension, and 52 as maximum length.
- Conv2D Layer: In this layer, the reshaped embedding matrix that is generated from the previous layer and the feature detector which is a randomly initialized matrix, are aligned together, then the feature detector is moved over the selected patch of the embedding matrix to perform element-wise multiplication, and finally, an element of the feature map is obtained through the convolutional operation and this feature map is the output matrix. The parameters of this layer are as follows: the number of filters is 512, kernel size is 5, embedding dimension is 64 and the activation function being used is 'Relu'.
- MaxPool2D Layer: The objective of this layer is to select the most contributing features. To do this, first of all, the window size is selected. In the next step, a stride of size 1 is selected so that the movement of the window over these feature maps is controlled using this and as a result, the maximum value is selected and inserted into the pooled feature map. It is the output matrix generated by the max-pooling operation.
- Flatten Layer: Multi-dimensional output created in the previous layer is converted into a single-dimensional form in this layer.
- Dropout Layer: To avoid overfitting in the network, a dropout layer is added where the dropout value is settled as 0.6 after executing some experiments.
- Output Dense Layer: The final output layer is a fully-connected dense layer that has some units that are equal to the total number of classes in our dataset, which is 21. Softmax is used as the activation function here and it computes the probability of the classes. 'Adam' optimizer is used in this layer.