2020 23rd International Conference on Computer and Information Technology (ICCIT), 19-21 December, 2020.

Sentiment analysis in Bengali via transfer learning using multi-lingual BERT

Khondoker Ittehadul Islam
Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet, Bangladesh
shanislam6@gmail.com

Md Saiful Islam
Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet, Bangladesh
saiful-cse@sust.edu

Md Ruhul Amin
Computer and Information Science
Fordham University
New York, USA
mamin17@fordham.edu

2020 23rd International Conference on Computer and Information Technology (ICCIT) | 978-1-6654-2244-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCIT51783.2020.9392653

***Abstract*—Sentiment analysis (SA) in Bengali is challenging due to this Indo-Aryan language's highly inflected properties with more than 160 different inflected forms for verbs and 36 different forms for noun and 24 different forms for pronouns. The lack of standard labeled datasets in the Bengali domain makes the task of SA even harder. In this paper, we present manually tagged 2-class and 3-class SA datasets in Bengali. We also demonstrate that the multi-lingual BERT model with relevant extensions can be trained via the approach of transfer learning over those novel datasets to improve the state-of-the-art performance in sentiment classification tasks. This deep learning model achieves an accuracy of 71% for 2-class sentiment classification compared to the current state-of-the-art accuracy of 68%. We also present the very first Bengali SA classifier for the 3-class manually tagged dataset, and our proposed model achieves an accuracy of 60%. We further use this model to analyze the sentiment of public comments in the online daily newspaper. Our analysis shows that people post negative comments for political or sports news more often, while the religious article comments represent positive sentiment. The dataset and code is publicly available [1].**

***Index Terms*—Sentiment Analysis, CNN, LSTM, BERT, GRU, fasttext, word2vec, SA, Bangla, Bengali**

I. Introduction

Sentiment classification is the task of analyzing a piece of text to predict the orientation of the attitude towards an event or opinion. The sentiment of a text can be either positive or negative. Sometimes, a neutral perspective is also considered for classification. SA has many different applications, such as reducing the early age suicide rate by identifying cyberbullying [1], discouraging unwarranted activities towards a particular community through hate-speech detection [2], and monitoring public response towards a proposed government bill [3] among many others.

The task of SA has achieved superior improvement in other languages, i.e. English - about 97.1% accuracy for 2-class [4] and 91.4% accuracy for 3-class SA [5]. But only a few research works have been published for the SA in Bengali. This is because we lack quality datasets in Bengali for training a computation model for the sentiment classification. However, in the last few years, we have seen the rise of Internet users in the Bengali domain mostly due to the development of wireless network infrastructure throughout South East Asia. This resulted in a massive increase in the total number of online social network users as well as newspaper readers. So it became comparatively easier to collect the public comments posted online on the Bengali news websites.

TABLE I: SA of public comment published in the online newspaper. We collected 334 comments for each of the politics, sports, and religion categories. We only collected one comment from a randomly selected news article. In the table, we present the percentage of the total comments classified into three different sentiment classes.

	Negative	Neutral	Positive
Politics	66%	24%	10%
Sports	52%	38%	10%
Religion	42%	8%	50%

Thus we created two SA datasets for 2-class and 3-class SA in Bengali and trained a multi-lingual BERT model via transfer learning approach for sentiment classification in Bengali, referred as $BERT_{BSA}$ in this paper. $BERT_{BSA}$ achieves an accuracy of 71% for the 2-class and 60% for the 3-class manually tagged dataset. We further use this model to analyze the sentiment of 1,002 public comments collected from the online daily newspaper. Table I shows that in general, sentiment in public comments is positive for religious news articles, while that is negative for political or sports news articles. In this paper, we present the following contributions:

- We created two datasets for SA in Bengali and made it public for further research work. We discuss the methodology we used to create the datasets in the Section III.
- We introduce a deep learning model for SA in Bengali, $BERT_{BSA}$, that performs better compared to other existing models that are trained with word2vec or fastText embedding. We discuss the model and in Section IV.
- We evaluate $BERT_{BSA}$ and compare it to other models trained with Word2Vec and fastText embeddigns using the 2-class and 3-class Bengali SA datasets. We discuss the results in the Section V.
- We conduct experiments to investigate application level use of Bengali SA on newspaper comments in three

[1]https://github.com/KhondokerIslam/Bengali_Sentiment

978-1-6654-2244-4/20/$31.00 ©2020 IEEE