International Conference on Bangla Speech and Language Processing (ICBSLP), 27-28 September 2019

Cricket Sentiment Analysis from Bangla Text Using Recurrent Neural Network with Long Short Term Memory Model

Md. Ferdous Wahid
Dept. of Electrical and Electronic Engineering
Hajee Mohammad Danesh Science and Technology University,5200
Dinajpur, Bangladesh
mfwahid26@gmail.com

Md. Jahid Hasan
Dept. of Electrical and Electronic Engineering
Hajee Mohammad Danesh Science and Technology University,5200
Dinajpur, Bangladesh
jahidnoyon36@gmail.com

Md. Shahin Alom
Dept. of Electrical and Electronic Engineering
Hajee Mohammad Danesh Science and Technology University,5200
Dinajpur, Bangladesh
ashahin200@gmail.com

2019 International Conference on Bangla Speech and Language Processing (ICBSLP) 978-1-7281-5241-7/20/$31.00 ©2020 IEEE 10.1109/ICBSLP47725.2019.201500

Abstract— **Nowadays, people used to express their feelings, thoughts, suggestions and opinions on different social platform and video sharing media. Many discussions are made on Twitter, Facebook and many respective forums on sports especially cricket and football. The opinion may express criticism in different manner, notation that may comprise different polarity like positive, negative or neutral and it is a challenging task even for human to understand the sentiment of each opinion as well as time consuming. This problem can be solved by analyzing sentiment in respective comments through natural language processing (NLP). Along with the success of many deep learning domains, Recurrent Neural Network (RNN) with Long-Short-Term-Memory (LSTM) is popularly used in NLP task like sentiment analysis. We have prepared a dataset about cricket comment in Bangla text of real people sentiments in three categories i.e. positive, negative and neutral and processed it by removing unnecessary words from the dataset. Then we have used word embedding method for vectorization of each word and for long term dependencies we used LSTM. The accuracy of this approach has given 95% that beyond the accuracy of previous all method.**

Keywords— ***sentiment analysis, natural language processing, deep learning, word embedding, RNN, LSTM.***

I. INTRODUCTION

In present era, people across the globe express their opinions or feelings through social media and web on different entities such as events, products, social issues, organizations etc. Hence, in every instant massive amount of text data are generated on various entities over the Internet. By analyzing these data business organizations can understand the sentiment of people about their products and can find new opportunities, government can understand people perception about election and can manage their reputation, event organizer can understand people expectation on public events and so on. Thus, it is a high need to epitomize the unstructured data created by people over the social media and extract relevant insights in order to understand people thoughts. Therefore, Sentiment Analysis has become a major point of focus in the field of NLP which extract contextual mining from text data that conveys emotions, sentiments or opinions of an individual.

In recent times, Cricket has gained uttermost popularity in Bangladesh. So, people have diversified emotions for this sport. They like to express their opinions, emotions regarding to this sport most often through social network in Bangla language. By processing these reviews, it is possible to understand the sentiment of people for cricket. However, very few attempts were taken for sentiment analysis from Bangla text because of the unavailability of well-structured resources in Bangla language processing. Hence, Cricket sentiment analysis on Bangla text from real people sentiments for cricket has become an exciting field for us. Nowadays, Deep learning technique is widely used to analyze sentiment of text and has proven to be an effective tool in terms of accuracy as it considered past and future word with respect to target word for text classification. Thus, we are very much influenced to classify cricket sentiment from Bangla text using deep learning technique.

In this research, Recurrent Neural Network with LSTM model has been proposed to identify cricket sentiment from Bangla texts. We have collected real people sentiments about cricket from different social media and news portal and categorized into positive, negative or neutral. Then, vectorization of each word was performed by word embedding method and LSTM was used to achieve long term dependencies. Finally, the accuracy of 95% has attained in cricket sentiment analysis using the proposed model.

II. RELATED WORK

Sentiment analysis from Bangla text has become major point of focus in NLP for researchers with the increasing use of social media. Hasan et al. [1] proposed a model utilizing LSTM with binary cross-entropy and categorical cross-entropy loss function for sentiment analysis of Bangla and Romanized Bangla Text (BRBT). Sharfuddin et al. [2] developed an approach combining deep RNN with bidirectional LSTM to classify sentiment of Bengali text which achieved 85.67% accuracy on a dataset containing 10000 comments of Facebook status. Baktha et al. [3] have explored RNN architectures on three dataset and obtained best accuracy from Gated Recurrent Units (GRU) for sentiment analysis. Tripto et al. [4] suggested deep learning based models for detecting multilabel sentiment and emotions from Bengali YouTube comments. They used Convolutional Neural Network (CNN), LSTM, Support Vector Machine (SVM) and Naïve Bayes (NB) architectures to identify three (positive, negative, neutral) and five label (strongly positive, positive, neutral, negative, strongly negative) sentiment as well as emotions where they considered SVM and NB as their baseline methods. Term Frequency Inverse Document Frequency (TF-IDF) with n-gram tokens has been used to extract set of features from respective sentence. They got 65.97% and 54.24% accuracy for three and five labels

978-1-7281-5241-7/19/$31.00 ©2019 IEEE

sentiments respectively. Sentiment polarity detection approach has been investigated on Bengali tweet dataset by Sarkar et al. [5] by applying multinomial NB and SVM. A character level supervised RNN approach was used to classify Bengali sentiment which is categorized as positive, negative and neutral [6]. An Aspect-Based Sentiment Analysis has been evaluated on Cricket comment in Bangla text in [7] where best accuracy has been found 71% using SVM classifier in their ABSA dataset. Shamsul et al. [8] employed SVM, Decision Tree and Multinomial NB for sentiment analysis on Bangladesh cricket comments. They got best accuracy using SVM which is 74%. So, we were influenced to develop a model for analyzing sentiment of real people by utilizing RNN with LSTM model.

III. Dataset Preparation

The preparation phases of the dataset were divided into two parts, i.e. A. Gathering of Bangla comments and B. Pre-processing of Bangla comments.

A. Gathering of Bangla comments

We have collected a dataset named 'ABSA' [7] that contains cricket related comments for cricket sentiment analysis from Bangla texts. The dataset comprised of 2979 data with 5 columns where we have selected only two columns, i.e. the comment column and the target column. The target column contains 3 classes including positive, negative and neutral. But, proposed LSTM-RNN model may create high variance problem on small dataset. To overcome this issue, we supplemented more data with existing ABSA datasets. The extended data were picked from various online resources like Facebook, YouTube, Prothom-Alo, BBC Bangla, Bdnews24.com and labelled them manually. Then, the extended ABSA dataset contains total 10000 comments where 8000 comments is separated for training and the rest of the comments is used for testing purpose.

B. Preprocessing of Bangla comments

Data pre-processing plays significant role in natural language processing (NLP) as the real-world data are messy and often contains error, unnecessary information and duplication. So, in order to generate good analytics results, all punctuation, unimportant words are removed, stemmed to their roots, all missing values are replaced with some values, case of text are replaced into a single one and mostly depending upon the requirement of the application. Therefore, we process our data step by step as it doesn't carry much weightage in context of the text. All preprocessed step is illustrated below.

1) Stopwords Removal: Stopwords refers to the most common words in a language. But these words have no impact on analysis sentiment of a sentence. The most common words such as এবং, এবার, এ, এটা, কী, হয়, র, পর, ওরা, কে, কেউ, ইত্যাদি have no impact on sentiment analysis using proposed model. But there some words such as না, নাই, নেই, নয় have important impact on negative sentiment and some words such as হ্যা, স্পষ্ট, করে, কাজ, কাজে have also important impact on positive sentiment. So, we list these positive and negative sentiment impact words from stopwords list as a whitelist. Then programmatically we have removed all stopwords from our dataset. We use two resources as a benchmark for removing Bangla stopwords.

2) Text Process: Text processing is of great importance in NLP task. The unwanted text such as Links, URLs, user tags and mention from comments, hash-tags, punctuation marks have no impact on sentiment analysis. Therefore, we remove these to give annotators an unbiased text. Only contents have given higher priority to make a decision based on three criteria positive, negative and neutral.

3) Name Process: Name process is another text data compression technique where mainly all proper and common noun of Bangla is substitute by common words. Example-

১. তামিম আজকে খুব সুন্দর ব্যাটিং করেছে।
২. লিটন দাশকে আজকে দলে নেয়া উচিত ছিল।

Here তামিম and লিটন are two different words but same context and also contribute similar impact on sentiment analysis using proposed LSTM-RNN model. So, we replace all proper noun with common word which does not affect the accuracy of the model but compress the dataset and makes dataset more robust and better for classification accuracy. Thus, we replace all country name such as (বাংলাদেশ, ভারত, পাকিস্তান, অস্ট্রেলিয়া) by a word "দেশ" and all players name including different spelling and nick name such as (তামিম, তামিম ইকবাল, মাশরাফি, ম্যাশ, সাকিব, বিরাট কোহলি, মুশফিকুর রহিম, মুশফিক) by word "সে" .

4) Manual validation: In manual validation, each extended data sample was manually annotated by three annotators into one of three categories: (i) positive (ii) negative and (iii) neutral. Each annotator validated the data without knowing decisions made by other. This ensures that the validations were unbiased and personal. Furthermore, elongated words often contain more sentiment information for multiclass categorizations. For example, "বাহ্হ্হহ অনেক ভালো!!" certainly provides more positive feelings. Therefore, instead of applying lemmatization we had kept elongated words.

IV. Methodology

Recurrent Neural Network (RNN) performed well enough on sequential input data like speech, music, text, name entity recognition etc. than convolutional neural network (CNN) by taking into consideration the current input as well as previous input. However, in long term dependencies it not generally works well due to vanishing and exploiting gradient problem [9]. Hence in order to learn long term dependencies, LSTM is simply added to the process input which enables RNN to remember their inputs over a long period of time. Here, this LSTM layer is fed with proper numerical vector representation of each word which is generated based on the word embedding. To generate word embedding, we have employed word2vec algorithm that formulates matrix of weight from text corpus. Finally, the output of LSTM layer is sent to fully-connected softmax layer to analysis sentiment of a comment.

A. Word Embedding

The preprocessed text data is tokenized first in order to split a sequence of sentence into smaller parts such as words. Then we implemented skip gram [10] model of word2vec algorithm for representing proper numerical vector of each splitting word that can evaluate the similarity between words and also placed close together in the vector space which is computationally effective for learning word embedding than one-hot vector representation. The output of the word2vec model is called an embedding matrix. The complete word embedding process is shown in Fig. 1.

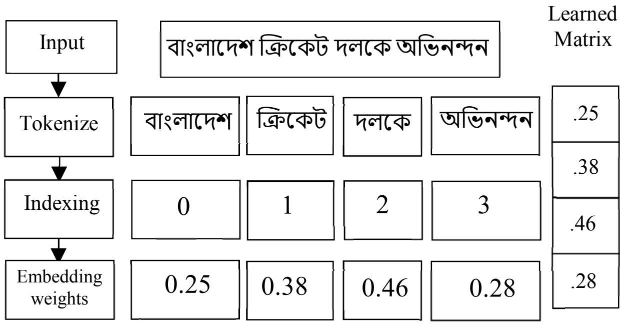

Fig. 1. Complete word embedding process

B. Model Architecture for sentiment analysis

The proposed architecture built to classify the sentiment of the cricket comments, consists of 4 layers. These layers are

1. Input layer (length 1)
2. Embedding layer (Output of length)
3. LSTM layer (100 neurons)
4. Output layer (103 neurons) (including 100 dense layers and 3 sigmoid layers)

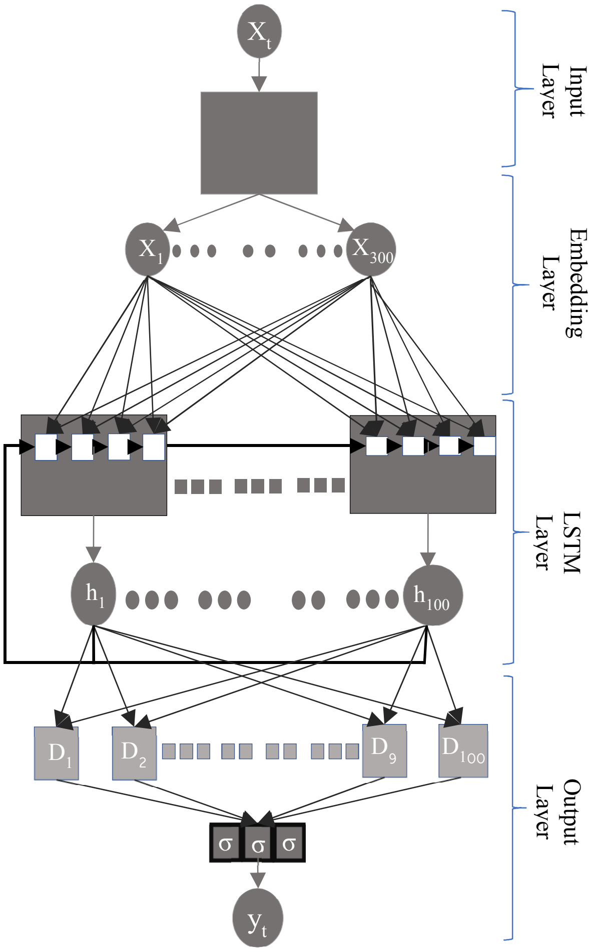

Fig. 2. LSTM network for cricket comments sentiment classification

A length of 42 words is used as a maximum word length of a comment and zero padding are added to the right of the comment when it is shorter than 42 words. The model takes input of 42 integers or vector of words, where each integer represents a word. So, there are 42-time steps, at each time step one word is given to the model. Then, the word is entered into the embedding layer with one neuron. The embedding layer transform the word into a numerical vector representation of length 300 (embedding size). The embedding weights are initially set to very small value and will update these weights using back-propagation during training. This way 300 featured value are created. Then, the output of embedding layer is fed into an LSTM layer with 100 neurons and each of the features value is multiplied by a weight of each LSTM cell, where LSTM cell contains four gates for long term dependencies memorization. Next those 300 weighted features and the output of the previous time step (output values from 100 neurons) is also used as an input for the LSTM cells. Finally, the weighted sum of dense layer outputs is taken as an input of softmax activation function where we predict the probability of cricket comments as positive, negative and neutral. The complete architecture is shown in Fig. 2.

V. Result and Discussion

We have used LSTM layer to construct and train many-to-one RNN architecture. The architecture takes sequence of words (sentence) as input and provides sentiment value (positive, negative or neutral) as output. We used data from the prepared dataset that contains 10,000 sentences in the ratio of 80:20 for training and test phase of proposed RNN-LSTM architecture respectively. During training phase, learning rate of the proposed architecture was set to a small value of 0.001 and different combinations of epochs and batch-size were used for attaining high prediction performance with minimum training time. The architecture attained best training accuracy after 15 epochs with batch size 30. Finally, the performance of the proposed architecture was evaluated on the test dataset where it obtained 95% prediction accuracy. The proposed model accuracy and loss graph was recorded while we train the model. Fig. 3 and Fig. 4 shows the model accuracy and loss graph respectively.

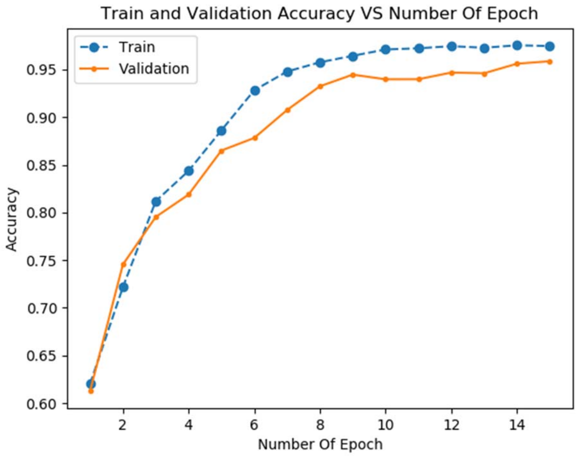

Fig. 3. Optimal accuracy of proposed model

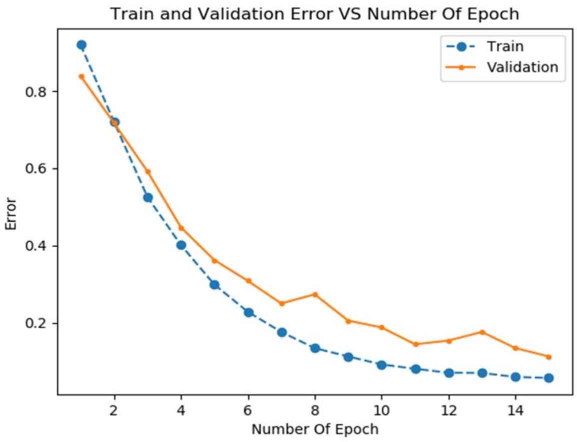

Fig. 4. Optimal loss of proposed model

Fig. 3 and Fig. 4 shows that the train and validation accuracy is being increased as well as is train and validation loss is being reduced respectively with respect to increasing the number of epochs. Example of some classification result is shown in TABLE I. The misclassification result is highlighted in the TABLE I using bold italic font.

TABLE I. SOME TEST RESULT

Sentence	Predict	Actual
বাংলাদেশ জিতবে ইনশাআল্লাহ।	positive	positive
টেষ্ট ক্রিকেটে রান আউট খুবই দুঃখজনক।	negative	negative
শাহারিয়ার নাফিস কে ও ফেরানো হোক।	positive	positive
ওরা ২০০ করছে তোমরা ১০০ করতে পারবে না	negative	negative
টস হারছে মানে, হারার সম্ভবনাই বেশী।	***neutral***	***negative***
মোসাদ্দেক, সাব্বির থেকে ভাল।	positive	positive
সেমিতে আমাদের প্রতিপক্ষ ভারত (প্রায় নিশ্চিত)।	neutral	neutral

Now a comparative study of classification accuracy among several models including proposed LSTM-RNN model is presented in Table II. It shows that the proposed architecture performs better than any other model of Bangla sentiment analysis, as RNN-LSTM model has a great competency to capture contextual information in more fine-grained way.

TABLE II. ACCURACY MATRIX AMONG DIFFERRENT BANGLA NLP TASK BASED ON MODEL PERFORMANCE

Ref No.	Dataset	Model	Prediction Accuracy
[7]	ABSA	SVM	71%
[8]	ABSA_EXTENDED	SVM	73.49%
Proposed method	ABSA_EXTENDED	LSTM	95%

VI. CONCLUSIONS

In this paper we present an approach to analyze sentiment of cricket comments in Bangla text. This model consists of a deep learning variant named RNN. For remembering the recurrent property and contextual meaning of a sentence we have used LSTM that makes the model very fruitful and produces a prediction result about 95%. Spell-checking and stemming is not included in preprocessing section of our collected dataset. In future, we will include more preprocessing steps along with these two in order to improve the structure of our dataset. We also plan to increase target class to make an accurate NLP model within this problem domain.

REFERENCES

[1] A. Hassan, M. Amin, A. Azad and N. Mohammed, “Sentiment analysis on bangla and romanized bangla text using deep recurrent models”, 2016 International Workshop on Computational Intelligence (IWCI), 2016.

[2] A. Aziz Sharfuddin, M. Nafis Tihami and M. Saiful Islam, “A Deep Recurrent Neural Network with BiLSTM model for Sentiment Classification”, 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018.

[3] K. Baktha and B. K. Tripathy, “Investigation of recurrent neural networks in the field of sentiment analysis”, 2017 International Conference on Communication and Signal Processing (ICCSP), 2017.

[4] N. I. Tripto and M. E. Ali, “Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments”, 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018.

[5] K. Sarkar and M. Bhowmick, “Sentiment polarity detection in bengali tweets using multinomial Naïve Bayes and support vector machines”, 2017 IEEE Calcutta Conference (CALCON), 2017.

[6] M. S. Haydar, M. A. Helal, and S. A. Hossain, “Sentiment Extraction From Bangla Text : A Character Level Supervised Recurrent Neural Network Approach”, 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018.

[7] M. Rahman and K. E.Dey, “ Datasets for Aspect-Based Sentiment Analysis in Bangla and Its Baseline Evaluation". Data, vol 3, issue 2, 2018.

[8] S. Arafin Mahtab, N. Islam and M. Mahfuzur Rahaman, “Sentiment Analysis on Bangladesh Cricket with Support Vector Machine”, 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, pp. 1-4, 2018.

[9] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, “Distributed Representations of Words and Phrases and their Compositionality”, NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems, vol 2, pp. 3111-3119, 2013.