3A comparative Study on Sentiment Analysis for Bengali Language Based on LSTM in F-Commerce

Conference Paper · October 2022

DOI: 10.1109/ICCCNT54827.2022.9984528

5 authors, including:

Mahafozur Rahman Rudro
Daffodil International University
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Md Faysal Wahid
Daffodil International University
1 PUBLICATION **1** CITATION
SEE PROFILE

Md ARIFUL Islam
Khulna University
6 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Nushrat Jahan Ria
Daffodil International University
19 PUBLICATIONS **49** CITATIONS
SEE PROFILE

A comparative Study on Sentiment Analysis for Bengali Language Based on LSTM in F-Commerce

Mahafozur Rahman
Department of CSE
Daffodil International University
Dhaka,Bangladesh
mahafozur15-2955@diu.edu.bd

Md Faysal Wahid
Department of CSE
Daffodil International University
Dhaka,Bangladesh
faysal15-2978@diu.edu.bd

MD ARIFUL ISLAM
Department of CSE
Daffodil International University
Dhaka,Bangladesh
ariful15-8406@diu.edu.bd

Nushrat Jahan Ria
Department of CSE
Daffodil International University
Dhaka,Bangladesh
nushratria.cse@diu.edu.bd

Dr. Sheak Rashed Haider Noori
Department of CSE
Daffodil International University
Dhaka,Bangladesh
drnoori@daffodilvarsity.edu.bd

***Abstract*—LSTMs are a type of RNN cell that regulates incoming and outgoing information using text-processing-specific gates. Online retailers, as we all know, have been extremely popular in recent years. Many social media sites, such as Facebook and Twitter, are increasingly used by online shops. As a result, market monitoring is critical. We can monitor user reviews and satisfaction levels using machine learning and deep learning methods. Many projects have been completed in various languages in the past. We use LSTM to work with Bangla text reviews in our research. We acquire an accuracy of 88 percent with this model.**

***Index Terms*—LSTM, RNN, merchants, machine learning, deep learning**

I. INTRODUCTION

These days, social media plays a significant role in our daily lives. Online social media content is widely available in a range of formats, such as blog posts, comments, and tweets [10]. Along with its numerous other uses, social networking is now a significant e-commerce platform. Large e-commerce companies have also expanded into social networking. According to Google, Bangladesh has 44.7 million Facebook users. Facebook commerce, or "f-commerce," is a relatively new online business venture that makes use of the social network as a virtual marketplace for advertising and conducting business transactions [3]. F-Commerce is being adopted by an increasing number of businesses. Numerous prospects for internet commerce are being created by it. Millions of money is thought to change hands each year in this industry. Bangladesh, one of the world's economies that are expanding quickly, has embraced digitization in many areas, including F-Commerce.Over 300,000 F-commerce pages are active in Bangladesh, but there are only roughly 2000 dedicated eCommerce sites, according to Google. Thus, it becomes necessary to continuously check the platform's quality.

This exploratory study aims to track the spread of online commerce and its effects. Previously, there has been a lot of work on vectorization in the field of sentiment analysis. In those cases, classifiers like logistic regression, random forest have been used. RNN is a popular method for sentiment analysis. RNNs were created in a way that allows them to capture time series and sequential data. In an RNN, we multiply by the weights assigned to the output and input of the previous state, respectively. Then, in order to obtain the new state, we pass them on to the Tanh function. Now, we multiply the new state by a Tanh function output to obtain the output vector. RNN works well with smaller datasets without much complication on the network but when it goes to the largest dataset performance of RNN are decreased. Two main reasons are the Vanishing Gradient Formula and the other one is Exploding Gradient Formula. Long short-term memory (LSTM) networks are a type of RNN that uses special units in addition to standard units. LSTM recovers the RNN vanishing or exploding gradient problem. It is able to recognize to keep information on the long memory and some information on the short memory. so that it can handle a huge amount of data. In this paper, LSTM is used for sentiment analysis and compared the performance with Random Forest Classifier, Naive Bayes Classifier, and SVM.

II. LITERATURE REVIEW

Many scholars have recently focused on sentiment analysis [10].The term polarity detection has appeared in this continuation. Kamal et al. [10] demonstrated an LSTM-based algorithm for Bengali tweet sentiment polarity identification. Numerous studies have demonstrated the efficacy of deep learning-based approaches for identifying sentiment polarity [7] [15]. Sanguansat et al. [9] showed Paragraph2Vec is better than TF and TF-IDF for Thai text processing for monitoring retail business, banking, and telecommunication in Thailand. According to the paper, Paragraph2Vec with Logistic Regression obtains 85.12 percent accuracy, which is better than any other combination. As we work with Bangla reviews it is necessary to analyze the performance of the Bangla text model. In the recent past, many works have been done to analyze Bangla text.

Ritu et al. [8] discussed the effectiveness of three three-word embedding models, notably word2vec in word2vec, Tensorflow from the Gensim package, and FastText model, on a Bangla text with 5,21,391 unique words.

Sharma et al. [11] detected Word2Vec, TF-IDF, and conventional CNN architecture were combined to extract features from text documents in Bangla humorous news that was distributed in online news portals and social media with an accuracy of more than 96 percent.

Tuhin et al. [14]work with Bangla text for detecting six types of emotions using the Naïve Bayes Classification Algorithm and Topical approach containing accuracy 70 percent for Naïve Bayes and 90 percent for Topical approach.

Sumit et al. [13]showed 83.79 percent accuracy was achieved using the Word2vec SkipGram and Continuous Bag of Words word embedding techniques, along with a Word to Index model, for SA in the Bangla language. In relation to Bengali text summarization, Abujar et al. [1] have been discussed.

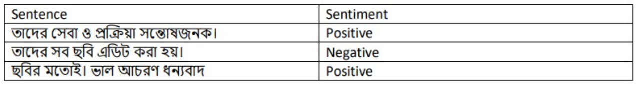
Alvi et al. [2] used CountVectorizer to extract the feature before applying Logistic Regression as a machine learning technique for the classification challenge. Sinha et al. [12] showed performance of word2vec as word embedding model. Long et al. [6]showed how to overcome the limitations of traditional machine learning-based sentiment analysis, they investigated the sentiment analysis of social media Chinese text using a mix of Bidirectional Long-Short Term Memory (BiLSTM) networks and a Multi-head Attention (MHAT) mechanism.. To extract emotion labels from psychiatric social literature, Jheng-Long Wu et al. [15] suggested using word embeddings, bidirectional long short-term memory (BiLSTM), and convolutional neural networks as part of a deep learning framework (CNN).

There has been a lot of vectorization work done in the past. Additionally, there is a problem with dataset size and no work has been done on binary classification. In this study, we work on binary text classification as well as dataset expansion.In this research, we attempt to construct a Bangla text processing model based on LSTM for social media market monitoring. We achieve this by gathering dataset from consumer feedback on various Facebook pages. Many recent studies have used LSTMs for sentiment analysis because they sidestep the long-term dependency problem and we get more accuracy by using this.

III. Methodology

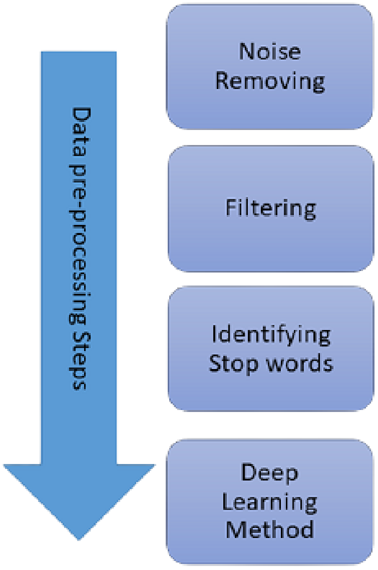
A. Dataset

We have collected data from various Facebook pages of the various online shop. We gathered feedback from those pages 'review sections. We gathered 2,000 reviews from various pages, weighing good and negative feelings. There were a few null values that we removed. There are 1,430 good ratings and 520 negative reviews for this product.



Sentence	Sentiment
তাদের সেবা ও প্রক্রিয়া সন্তোষজনক।	Positive
তাদের সব ছবি এডিট করা হয়।	Negative
ছবির মতোই। ভাল আচরণ ধন্যবাদ	Positive

Fig. 1. Sample Dataset



B. Data Preprocessing

We processed our dataset after collecting data. We eliminated out noise, identified useless comments, and used stop words. Many emojis and symbols have been removed, including

$$[], -,, =, :, +, @, !, , , ; , /, ,), (,], , , (:,)$$

. After that, we used deep learning and comment classification methods.

Fig. 2. pre-processing steps

C. Experiments

We employed LSTM to analyse consumer reviews on the internet market. TF, TF-IDF, Word2vec, Naive Bayes, and SVM were previously utilized for online market analysis and Bangla text analysis referring [9] [8] [14]. However, because our research is focused on sentiment analysis, LSTM has recently been deployed, and we've discovered that it performs well in many sentiment analysis studies referring [10] [5] [4] . For LSTM at first, we tokenize our sentences and finally fit our model.

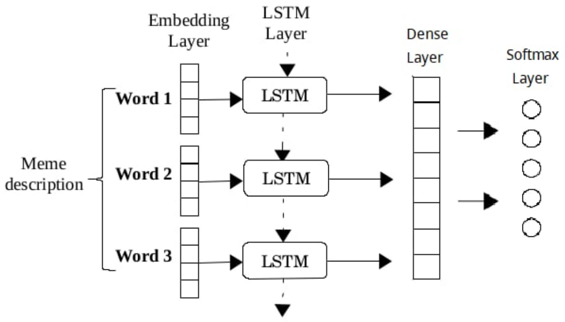


Fig. 3. LSTM Method

Figure 3 shows the general architecture and it was used to create our sentiment detection model based on LSTM. The output of the LSTM unit is passed through the softmax dense layer. It predicts the class of the input texts, as shown in Figure 3. Figure 4 depicts how the LSTM model works.

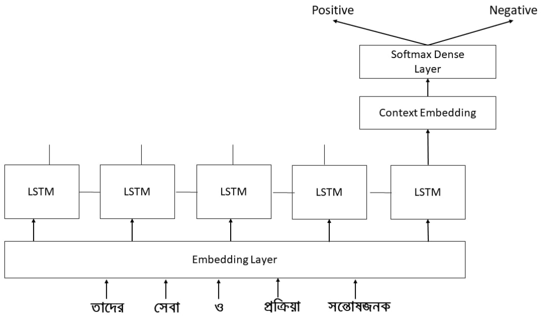


Fig. 4. LSTM over a sample text

D. LSTM Method

The long-term reliance problem is explicitly avoided using LSTMs. They don't have to work hard to remember things for long periods of time. Recurrent neural networks (RNNs) have a problem with long-term dependability that LSTM networks were created to solve. In contrast to more common feedforward neural networks, LSTMs have feedback connections. This feature allows LSTMs to handle whole data sequences without having to treat each data point in the series separately by retaining important information from earlier data in the sequence to aid in the processing of incoming data points. As a result, LSTMs are excellent at processing time series, text, and other sequential data types. An LSTM network may learn this pattern, which occurs every 12 periods of time It avoids the long-term reliance problem that other models have by not simply using the prior forecast but also keeping a larger context in mind. However, LSTMs become increasingly useful when patterns are separated by noticeably longer periods of time. It should be noted that this is a rather simple example.

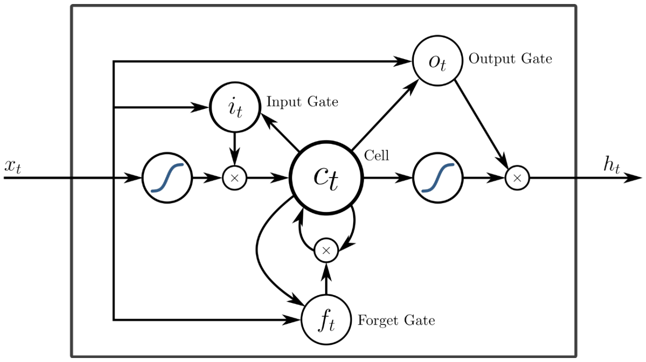


Fig. 5. LSTM Method

simple LSTM memory cell's framework. As shown, this architecture has three gates (ft, it, and ot) as well as a memory cell (ct).

Each LSTM cell's formula can be written as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (2)$$

$$i_t = \sigma(W_f x_t + U_i h_{t-1} + b_i) \quad (3)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (4)$$

The parameters for gates and cell states are W, U, and b. Each LSTM cell's forget, input, and output gates are specified by these three equations. in that order, eqs. 2-4. As depicted in Figure 3, the forget gate in an LSTM layer determines which prior information from the cell state is forgotten. The input gate determines or regulates the new information that is stored in the memory cell. The quantity of data that may be retrieved from the internal memory cell is controlled or limited by the output gate.

E. Naive Bayes Classifier

Naive Bayes is a machine learning approach. It handles enormous amounts of data. It does a fantastic job at NLP tasks like sentimental analysis. It's a categorization algorithm that is very simple and quick. A method for examining conditional probabilities is the Bayes rule. It gives you a convenient way to turn the situation around. The Bayes rule is the foundation of Bayesian classifiers. Given the evidence B, a conditional probable means that event A is likely to occur. We use notation:

$$P(A|B)$$

.It is the standard notation. It is the standard notation. When we only have the likelihood of the opposite result and of the two components separately, we can use the Bayes method to calculate this probability:

$$P(A|B) = P(A)P(B|A)/P(B)$$

.When estimating the chance of something based on examples of it happening, this restatement can be quite useful.

F. Random Forest Classifier

Random Forest is one of the most well-known machine learning algorithm. It is used in the supervised learning method. Random forest can be used for classification as well as regression problems in machine learning. It is based on ensemble learning, a technique for combining several classifiers to tackle a challenging problem and improve the performance of the model. " A Random Forest classifier combines several decision trees on various dataset subsets and then averages them to raise the anticipated accuracy of the dataset." Instead than depending on a single decision tree, the random forest gathers the predictions from each tree and predicts the ultimate result based on the majority votes of predictions.

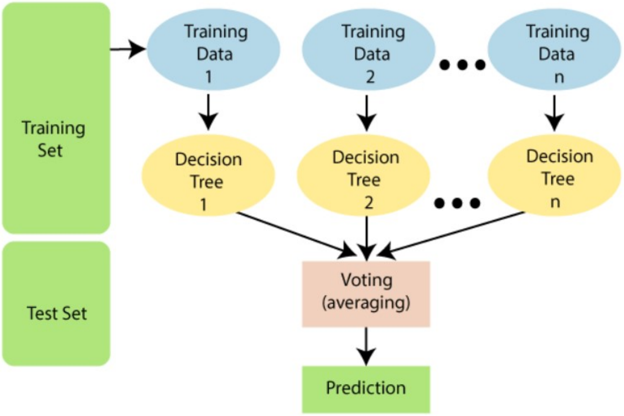


Fig. 6. Random Forest Classifier

G. Support Vector Machine Algorithm

The Support Vector Machine, or SVM is a common Supervised Learning technique. It is used for Classification and Regression issues. However, it is mostly utilized in Machine Learning for classification difficulties. The SVM method's goal is to identify the best line or decision boundary for categorizing n-dimensional space so that future data points can be quickly assigned to the appropriate category. The best possible chosen boundary is a hyperplane. SVM chooses the extreme points of the hyperplane. The most severe examples of this are called support vectors, and this process is called a support vector machine.

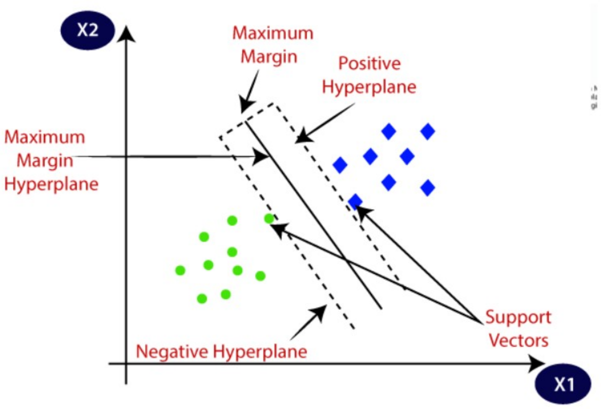
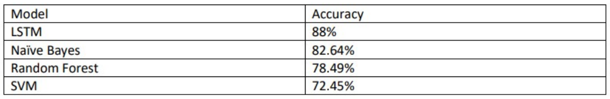


Fig. 7. SVM

IV. Result and Comparisons

For our Bangla text, we employed LSTM, which produced an accuracy of 88 percent. Then we applied three more machine learning classifiers: Naive Bayas (82.64 percent accuracy), Random Forest (78.49 percent accuracy), and SVM (72.45 percent accuracy).



Model	Accuracy
LSTM	88%
Naïve Bayes	82.64%
Random Forest	78.49%
SVM	72.45%

Comparison of the result is showed graphically in below.

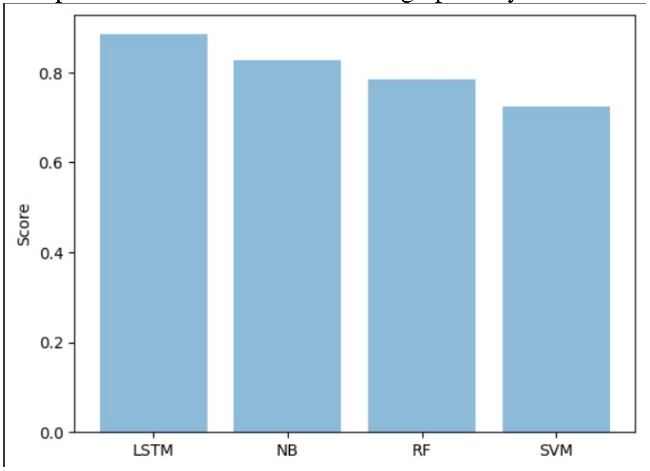


Fig. 8. Comparison of Results

V. Conclusion

The Bangla language sentiment analysis is proposed in this study.In the past, many works of sentiment analysis have been done in different languages. There is a lot of emphasis on vectorization. BERT model has been used for sentiment analysis in Bengali language. It is a feature of our platform that allows us to keep track of Bangladesh's social media market analysis. We're trying to show that LSTM is very useful in Bangla text processing based on experimental results. There have some limitations. The dataset we have collected is not balanced.In future, we can work on emoji detection by LSTM, increase our dataset and will work with unsupervised machine learning model.

References

[1] Sheikh Abujar, Abu Kaisar Mohammad Masum, Md Mohibullah, Syed Akhter Hossain, et al. An approach for bengali text summarization using word2vector. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE, 2019.

[2] Nasif Alvi and Kamrul Hasan Talukder. Sentiment analysis of bengali text using countvectorizer with logistic regression. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 01–05. IEEE, 2021.

[3] Mohammad Ali Ashraf, Mirza Mohammad Didarul Alam, and Lidia Alexa. Making decision with an alternative mind-set: Predicting entrepreneurial intention toward f-commerce in a cross-country context. *Journal of Retailing and Consumer Services*, 60:102475, 2021.

[4] Xin Huang, Wenbin Zhang, Xuejiao Tang, Mingli Zhang, Jayachander Surbiryala, Vasileios Iosifidis, Zhen Liu, and Ji Zhang. Lstm based sentiment analysis for cryptocurrency prediction. In *International Conference on Database Systems for Advanced Applications*, pages 617–621. Springer, 2021.

[5] Hamed Jelodar, Yongli Wang, Rita Orji, and Shucheng Huang. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2733–2742, 2020.

[6] Fei Long, Kai Zhou, and Weihua Ou. Sentiment analysis of text based on bidirectional lstm with multi-head attention. *IEEE Access*, 7:141960–141969, 2019.

[7] Prem Melville, Wojciech Gryc, and Richard D Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284, 2009.

[8] Zakia Sultana Ritu, Nafisa Nowshin, Md Mahadi Hasan Nahid, and Sabir Ismail. Performance analysis of different word embedding models on bangla language. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5. IEEE, 2018.

[9] Parinya Sanguansat. Paragraph2vec-based sentiment analysis on social media for business in thailand. In *2016 8th International Conference on Knowledge and Smart Technology (KST)*, pages 175–178. IEEE, 2016.

[10] Kamal Sarkar. Sentiment polarity detection in bengali tweets using lstm recurrent neural networks. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pages 1–6. IEEE, 2019.

[11] Arnab Sen Sharma, Maruf Ahmed Mridul, and Md Saiful Islam. Automatic detection of satire in bangla documents: A cnn approach based on hybrid feature extraction model. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5. IEEE, 2019.

[12] Manjira Sinha, Rakesh Dutta, and Tirthankar Dasgupta. Does word2vec encode human perception of similarityƒ a study in bangla. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4. IEEE, 2019.

[13] Sakhawat Hosain Sumit, Md Zakir Hossan, Tareq Al Muntasir, and Tanvir Sourov. Exploring word embedding for bangla sentiment analysis. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5. IEEE, 2018.

[14] Rashedul Amin Tuhin, Bechitra Kumar Paul, Faria Nawrine, Mahbuba Akter, and Amit Kumar Das. An automated system of sentiment analysis from bangla text using supervised learning techniques. In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pages 360–364. IEEE, 2019.

[15] Xingyou Wang, Weijie Jiang, and Zhiyong Luo. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 2428–2437, 2016.