IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Bangla Sentiment Analysis Using Artificial Intelligence

[1]Ismail Hossain Sadhin, [2]Elora Majumder Bandhan
[1]Computer Science, [2]Electrical & Electronics Engineering
[1,2]American International University-Bangladesh, Dhaka, Bangladesh

Abstract: For understanding and interpreting human emotions expressed in textual data as public opinion, reviews, or customer feedback, sentiment analysis is a vital section of natural language processing. By analyzing these sentiments, any organization can adopt proper strategies to enhance its effectiveness. Despite of being one of the most widely spoken languages globally, Bangla remains relatively underexplored in this domain. Rarely research has been conducted for Bangla Sentiment Analysis. In this work, sentiment analysis has been done based on the Bangla language, using a dual-method approach. Initially, sentiment expressions in Bangla text were analyzed with the sequential model. Following this, a comparative analysis is conducted using Support Vector Machines (SVM), that enhanced the accuracy of sentiment classification. The results reveal a notable improvement in accuracy and performance with the SVM approach.

IndexTerms **- Sentiment Analysis, LSTM, SVM, Bangla, NLP, Deep Learning.**

I. INTRODUCTION

Sentiment analysis is a natural language processing (NLP) technique that is used to analyze data and classify it in different polarity such as positive, negative, or neutral. This technique helps to explore information from different sources which saves a lot of time and labor. Sentiment analysis (SA) is one of the most dominating fields of research since there is a large portion of opinionated data on the Internet and other sources. These days, people express their reviews and opinions on social media sites, newspapers, blogs, etc. There is also forum discussion on specific posts, article or in the comments of these posts and articles, and even on products or services. There are many difficulties in detecting a class of sentiment such as subjectivity or opinion-based identification, when a phrase or text does have not any fundamental opinion word. Mining down this enormous data manually and identifying manifested opinions in a systematic manner can be both labor-intensive and could have countless errors. To solve this problem, diving through one of the most important branches of Natural Language Processing (NLP) is Sentiment Analysis because this area has large-scale problem domain. The fundamentally difficult task in this field is comprehending the intricate semantic structure of languages. Examining ambiguous statements in which positive words may indicate negative meanings or vice versa is a crucial case. As of now, beginning from advertising to client care in associations, online entertainment checking, political perspectives examination, and a lot more domains of living souls are outperformed by the give of sentiment analysis. According to a large number of studies, the most crucial aspects of business performance are the quality of the products or services provided and the contentment of customers [1]. To guarantee the organization's competitiveness, organizations should cautiously consider their customers' expectations and needs from the items or services they give. Likewise, they should well deal with their customers by making them satisfied to work with them [2]. The Bangla language, which is spoken by a large number of people, should be the focus of this important research. Bangladeshi individuals progressively participating in online exercises, for example, - associated with loved ones via web-based entertainment, offering their viewpoints and contemplations on well-known microblogging and informal communication locales, imparting insights and considerations through remarks on internet-based news entryways, doing internet shopping through internet-based commercial centers and other such applications. Therefore, the application of automated Sentiment analysis (SA) can play a vital role here. The paper will describe the sentiment analysis done over the Bangla Language using two algorithms and come up with a more accurate analysis.

II. RELATED WORK

Researchers are finding SA to be an intriguing issue in the age of social media and microblogging services. It appears that SA is conducted in a variety of languages, including Arabic, Chinese, English, and French. However, because of certain empirical and technological limitations, the extent of its advancement in the Bengali language is negligible. Yet there is little research work done on Bangla Sentiment Analysis. The deep Recurrent model's sentiment analysis has been done on Bangla & Romanized Bangla Text where a substantial textual dataset has been tested in Long Short-Term Memory (LSTM), using two types of loss functions – binary

cross-entropy and categorical cross-entropy, and also some experimental pre-trainings were conducted by using data from one validation to pre-train the other and vice versa [3]. Further Sentiment Analysis has been done on Bangla text using Supervised Machine Learning with an Extended Lexicon Dictionary that was developed for analyzing sentiments in Bangla. This LDD is developed by applying the concepts of normalization, tokenization, and stemming to a specific Bangla dataset [4]. Another work of sentiment analysis has been done on Bangla Newspaper using supervised machine Learning Algorithms where six classifiers have been used to come up with higher accuracy [5].

III. METHODOLOGY

The methodology employed in this study involved the application of two machine learning algorithms, namely Long Short-Term Memory (LSTM) and Support Vector Machine (SVM), for Bangla sentiment analysis. Initially, a dataset comprising 6652 samples was utilized for training the LSTM model. The LSTM architecture was configured with appropriate hyperparameters, including the number of layers, hidden units, and learning rate, followed by a rigorous training process to learn the intricate patterns of sentiment in Bangla text. Subsequently, the same dataset was subjected to sentiment analysis using SVM, with careful consideration given to parameter selection and optimization to ensure optimal performance. The SVM model was configured with a suitable kernel function and regularization parameter to effectively classify sentiment polarity. Performance evaluation of both models was conducted using standard sentiment analysis metrics, such as accuracy, precision, recall, and F1-score, to assess their effectiveness in accurately predicting sentiment in Bangla text. Our experimental results revealed that the SVM method outperformed LSTM in terms of all evaluation metrics, indicating its superior performance in Bangla sentiment analysis when trained on a dataset of 6652 samples. In the table of showing the polarity of the dataset for SVM shows among the 6652 samples, positive sentiments were 3037 samples and negative sentiments were 3615 samples. In Figure 1 it shows the graphical representation of several positive sentiments represented by 1 and negative sentiments represented by 0. This finding underscores the significance of selecting appropriate machine learning algorithms and optimizing their parameters for achieving accurate sentiment analysis results in the Bangla Language.

Table 1 Showing Polarity of the Test Dataset for SVM

Sl.	Number of Test Data in Each Polarity		
	Positive	Negative	Neutral
1	3037	3615	0

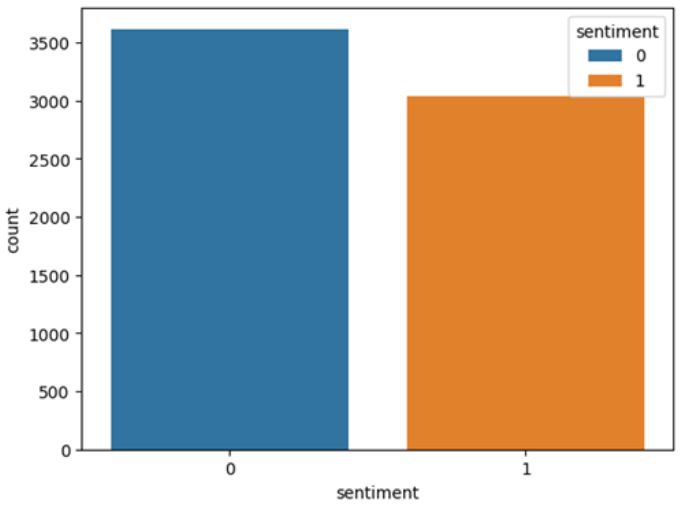

Figure 1 Positive & Negative Sentiment Count

3.1 Sequential Model

The idea behind the Sequential model methodology for sentiment analysis is to handle the info information on each piece in turn, sequentially, as a rule as per the pattern in which they show up. In the context of natural language processing, this refers to treating text as a sequence of words or tokens. Layers of neurons are used in the model, and long short-term memory networks (LSTMs) or recurrent neural networks (RNNs) are frequently used. These networks are particularly adept at recognizing temporal connections in sequence. The calculation looks further into the setting of each word it processes, which assists it with improving comprehension of the feeling being communicated. The model can catch nuances and connections in language structure that are helpful to this consecutive methodology, which is fundamental for accurately recognizing feelings in convoluted phrases. Using a sequential model to implant these temporal relations into user and product representations a document-level sentiment analysis, therefore it can be suggested that reviews' temporal relations could be useful for learning user and product embedding [6]. In response to the prediction error, the model's parameters are altered during training.

3.2 Support Vector Machine (SVM)

In Sentiment analysis, the Support Vector Machine (SVM) calculation works by making an ideal hyperplane that successfully isolates data of interest from various feeling classes. This algorithm has been used for various applications such as time series prediction, face recognition, biological data processing, etc. [7]. Concerning assessment examination, each message test is

tended to as a part vector, with features obtained from the words or articulations present. SVM intends to find a hyperplane that increases the difference between positive and negative feeling cases, taking into account improved speculation to hidden data. The estimation changes the data space into a higher-layered space, where the inclination classes become straightly detachable. Utilizing a piece capacity, SVM can capably manage non-straight associations in the data. During the readiness stage, the model sorts out some way to dispense feeling names in view of the ideal hyperplane, updating limits to achieve exact gathering. SVM is a strong choice for feeling examination tasks because of its ability to deal with complex choice limits and deal with highly layered information.

3.3 Deep Learning

Deep learning plays an important role in sentiment analysis, providing powerful tools to capture complex patterns and contextual nuances in text data. Neural network architectures, especially recurrent neural networks (RNNs) and long-term short-term memory (LSTMs) are commonly used for sentiment analysis tasks. These models excel at processing sequential data, allowing the contextual dependence of words in sentences to be taken into account. Another noteworthy architecture is the transformer, which has helped achieve state-of-the-art results in various natural language processing tasks, including sentiment analysis. These models use attention mechanisms to focus on different parts of the input text and efficiently capture long-range dependencies. Transfer learning, using pre-trained language models such as Bidirectional Encoders for Transformers (BERT), has also become popular, allowing emotion-specific datasets to be fine-tuned to improve performance. In general, deep learning techniques improve sentiment analysis by automatically learning complex hierarchical language representations, enabling more accurate and nuanced sentiment classification [8].

3.4 Recurrent Neural Work

For handling sequential data effectively by retaining data about previous inputs through recurrent connections, Recurrent Neural Network (RNNs) is designed in the classification of artificial neural networks. Unlike feedforward neural networks that process input data in a single pass, Recurrent Neural Network (RNNs) has loops within their architecture, which maintain internal states and process sequences of variable length. This unique architecture enables Recurrent Neural Network (RNNs) to capture temporal dependencies and context in sequential data, making them particularly suitable for applications such as speech recognition, natural language processing, and time series prediction. Besides, there are more applications of Recurrent Neural Network (RNNs) has studied such as statistical language modelling, finding outline of the paper, evaluation of different language models and so on [9]. However, traditional RNNs often limits the vanishing gradient problem, because of restricting their ability to effectively capture long-range dependencies. To solve this issue, more advanced RNN variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), have been developed, incorporating mechanisms to selectively retain or forget information over time. These enhancements have significantly improved the performance of Recurrent Neural Network (RNNs), making them a powerful tool for modeling sequential data in various domains.

IV. DATA PREPROCESSING

For accurate results, data preprocessing is crucial in sentiment analysis. After collecting data, these are preprocessed to create training data and testing data. We used SVM method for data preprocessing. The steps followed in the data preprocessed are described below-

4.1 Remove Punctuation

In different content, there could be several punctuation marks available for formatting or indicates the rest in a sentence or even use for decorative purpose. But these punctuations are not express any sentiment. So, these marks are removed for clear data to analyze. This step of removal streamlines text for NLP model, for focusing on the content not on formatting. It ensures a clear data so that the model extract the result from meaningful insights which helps to distinguish linguistic patterns and nuances in the text.

4.2 Remove Stop-Words

Removing the stop words is another step of text processing that involves in filtering out the common words or non-informative words such as “অবশ্য” , “অনেক” and "ছাড়াও" and so on. Besides in this process, the pronouns and articles also have been removed. The main purpose of removing these words is get the analysis outcome without any distortions. Because these words carry very minimal sentimental value. This process enhances the effectiveness of the tasks such as sentiment analysis or information retrieval. Techniques utilizing predefined lists or statistical methods are employed to systematically remove these functional words, streamlining subsequent text processing stages.n different content, there could be several punctuation marks available for formatting or indicates the rest in a sentence or even use for decorative purpose. But these punctuations are not express any sentiment. So, these marks are removed for clear data to analyze. This step of removal streamlines text for NLP model, for focusing on the content not on formatting. It ensures a clear data so that the model extract the result from meaningful insights which helps to distinguish linguistic patterns and nuances in the text.

4.3 Tokenization

Tokenization in sentiment analysis involves breaking down a text into individual words or sub-words, treating them as separate units or "tokens." This process facilitates analysis by providing a structured representation of the text. It's a fundamental step in preparing textual data for sentiment classification. For example, consider the sentence: “কারাগারে বিলাসী জীবন কাটছিল মুফতি হান্নানের।” [Mufti Hannan had a luxurious life in prison.], after tokenizing the sentence, it create a list of – [“কারাগারে” [prison], “বিলাসী” [luxurious], “জীবন” [life], “কাটছিল” [had], “মুফতি” [Mufti], “হান্নানের” [Hannan]]. After tokenizing, Normalizing is done by removing characters [“ , ”, “ . ” , “!”, “@”, “#”, “%”]. Each token is then analyzed to determine its sentiment contribution, aiding in the overall sentiment classification of the sentence.

4.4 Stemming

Stemming in sentiment analysis involves reducing words to their base or root form to capture the core meaning. This process helps in normalizing variations of words, enhancing the efficiency of sentiment analysis models. By applying stemming, variations of words are treated as the same, reducing complexity and improving the model's ability to recognize sentiment patterns. For example- "খারাপ মানুষের অত্যাচারের জন্য এই পৃথিবী ধ্বংস হচ্ছে না এই পৃথিবী ধ্বংস হচ্ছে ভাল মানুষ গুলোর নীরবতার জন্য" , in this sentence, the words "মানুষের" , "অত্যাচারের" , "নীরবতার" - convert info the root words respectively – "মানুষ" , "অত্যাচার" , "নীরবতা" .

4.5 Parts of Speech (POS) Tagger

Considering parts of speech in sentiment analysis involves analyzing the grammatical categories of words in a text to understand how they contribute to sentiment. This information can be valuable for extracting more nuanced sentiment from a sentence.

V. IMPLEMENTATION

In this study we implemented the model what we claimed before. Here we use our preprocess data and use our algorithm. Already you know that our applied algorithms are LSTM and SVM were using the SVM algorithm, we have got better and more accurate outcome.

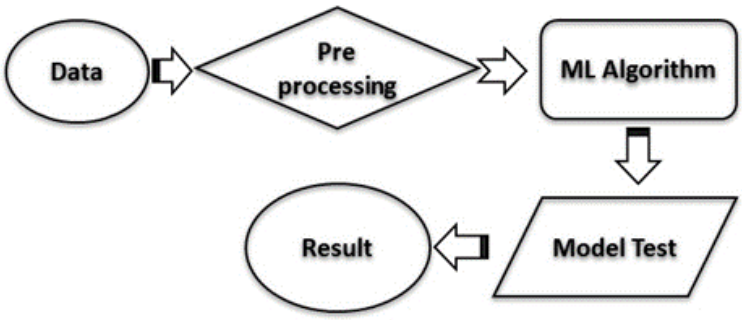

Figure 2 Steps of implementation of the study

In figure 2, it shows the steps where after getting data, it processed though different steps of data preprocessing method, during this process it followed a machine learning algorithm. For this study we chose the LSTM and SVM. Further the data has been tested and get the required result.

VI. RESULT & ANALYSIS

Within the framework of this research paper, our sentiment analysis methodology unfolds across a two-stage process. Firstly, we explored the complicated textual data by deploying sequential model. Building upon the insights gathered from the sequential model, we seamlessly transitioned into the second stage, implementing the Support Vector Machine (SVM) algorithm. The justification behind this dual-stage strategy was rooted in the collaborative combination of model capabilities. The sequential model, adept at capturing sequential dependencies, laid the foundation, while the SVM algorithm, with its robust classification capabilities, acted as a complementary force. This dual-stage methodology was strategically designed to influence the strengths of both models, resulting in a notable improvement in overall sentiment analysis accuracy.

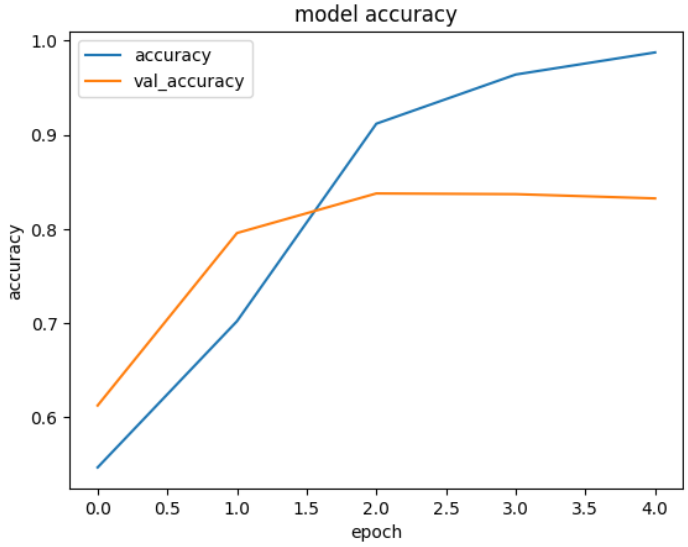

Figure 3 Model Accuracy Curve for SVM

In the figure 3 of model accuracy curve, after using SVM algorithm it explained for value accuracy between 0.6 to 0.9 the model will provide the accuracy up to 1 for different epoch. Another analysis could be done by the relation between the value loss and accuracy for test data-

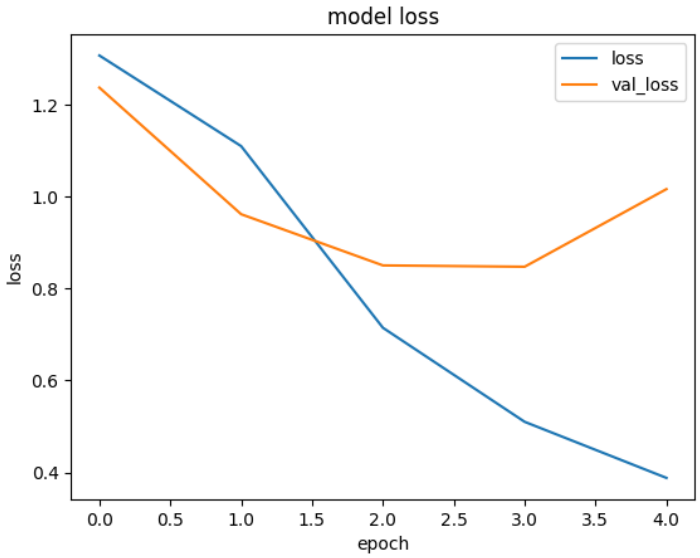

Figure 4 Model Loss Curve

In the figure 4 of model, it explains while test the data, for several epoch, loss of data varies from 0.8 to 1.2 with accuracy of 0.5 to 0.9. Again, for this value loss, the model loss varies from 0.4 to 1.2 and gave accuracy of 0.8.

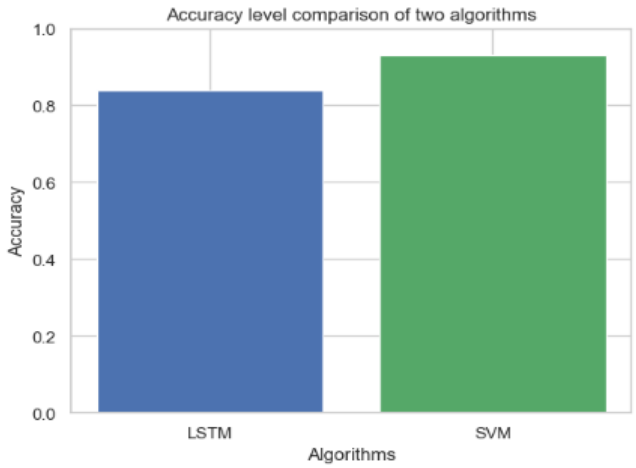

Figure 5 Accuracy comparison curve between LSTM & SVM

After the analysis, integration of a sequential model in our sentiment analysis framework has yielded a test accuracy of 0.857. After that, our methodology progresses with the introduction of the Support Vector Machine (SVM) algorithm, resulting in an impressive test accuracy of 0.936. In figure 5, accuracy of both algorithms has been compared. This strategic combination proceeded to achieve heightened accuracy in sentiment classification through a comprehensive and effective methodology.
Based on this analysis, a real time feedback could get from our Bangla sentiment analysis app. That explained with few examples in the following figures-

Figure 6 Bangla sentiment analysis application analyzed a positive expression

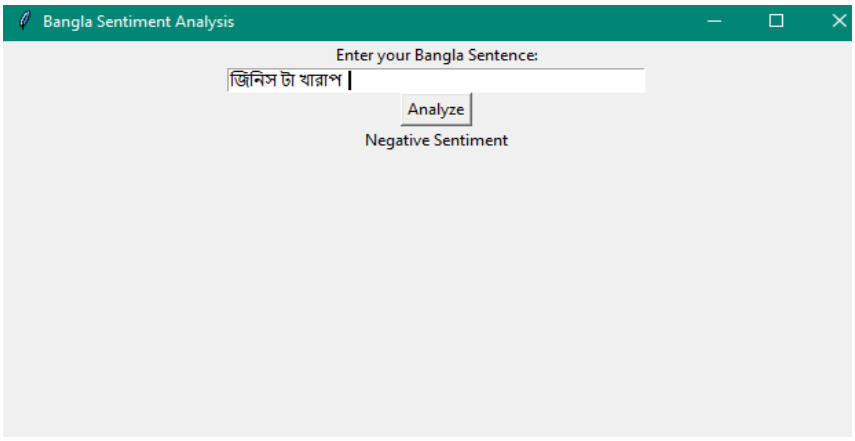

Figure 7 Bangla sentiment analysis application analyzed a Negative expression

The figure 6 shows that while a bangla sentence such as 'আমি অনেক ভালো। 'has been input in the application which is surely a positive impression, the analysis application provide the result as- 'Positive Sentiment'. Similarly, a negative impression such as 'জিনিসটা খারাপ।' results as 'Negative Sentiment' in the application has shown in figure 7. So that this application visually shows the accuracy of our model.

Receiver Operating characteristics (ROC) curve refers to the satistical accuracy of a model by ploting the Ture positive rate (TPR9 and False positive rate (FPR) in the axis. That shows the deviation between the accual value and predictated value and provide a curve for a specific model.

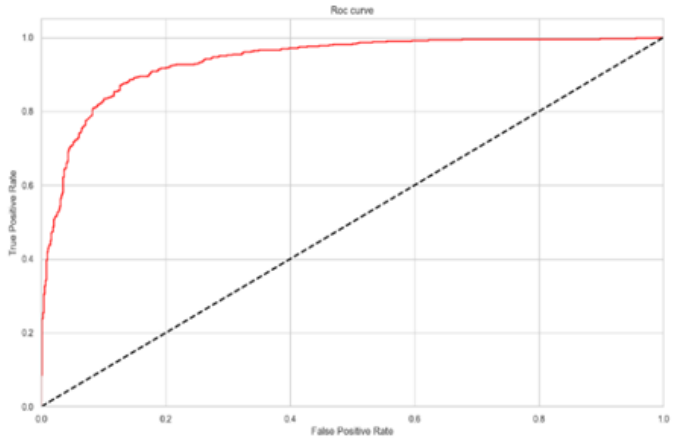

Figure 8 ROC Curve

The ROC curve showed in figure 8 of our model shows the threshold of TPR and FPR at 1 and around 0.5 respectively. So that, with this model this the best result accuracy can be obtain.

VII. CONCLUSION

In conclusion, sentiment analysis holds crucial significance in the Bangla language. This research, employing both a sequential model and SVM algorithm, emphasizes the importance of different approaches. The findings reveal that SVM outperforms the sequential model, emphasizing its efficacy in accurately discerning sentiments in Bangla text. This study contributes valuable insights for enhancing sentiment analysis in the Bangla language, paving the way for more effective language processing applications in this linguistic context.

REFERENCES

[1] A. engine, "Effect of customer satisfaction on company performance," Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, https://doi.org/10.11118/actaun201563031013 (accessed May 4, 2024).

[2] S. Ilias and M. F. Shamsudin, "Customer satisfaction and business growth," Journal of Undergraduate Social Science and Technology, http://abrn.asia/ojs/index.php/JUSST/article/view/60 (accessed May 4, 2024).

[3] A. Hassan, M. R. Amin, A. K. Azad, and N. Mohammed, "Sentiment Analysis on Bangla and romanized bangla text using Deep Recurrent models," 2016 International Workshop on Computational Intelligence (IWCI), Dec. 2016. doi:10.1109/iwci.2016.7860338

[4] N. R. Bhowmik, M. Arifuzzaman, M. R. H. Mondal, and M. S. Islam, "Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary," Natural Language Processing Research, https://doi.org/10.2991/nlpr.d.210316.001 (accessed May 4, 2024).

[5] S. J. Maisha, N. nafisa, and A. K. M. Masum, "Supervised machine learning algorithms for sentiment analysis of Bangla newspaper," International Journal of Innovative Computing, https://doi.org/10.11113/ijic.v11n2.321 (accessed May 4, 2024).

[6] T. Chen, R. Xu, Y. He, Y. Xia, and X. Wang, "Learning user and product distributed representations using a sequence model for sentiment analysis," Aston Research Explorer, https://research.aston.ac.uk/en/publications/learning-user-and-product-distributed-representations-using-a-seq (accessed May 4, 2024).

[7] T. Evgeniou and M. Pontil, "Support Vector Machines: Theory and applications," Machine Learning and Its Applications, pp. 249–257, 2001. doi:10.1007/3-540-44673-7_12

[8] M. S. Islam et al., "'challenges and future in deep learning for sentiment analysis: A Comprehensive Review and a proposed novel hybrid approach,'" Charles Sturt University Research Output, https://researchoutput.csu.edu.au/en/publications/challenges-and-future-in-deep-learning-for-sentiment-analysis-a-c (accessed May 4, 2024).

[9] Wim De Mulder, Steven Bethard, and M. F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," Computer Speech & Language, https://www.sciencedirect.com/science/article/pii/S088523081400093X (accessed May 4, 2024).