Bangla Financial Sentiment Classification Model Using Machine Learning

Alok Das
United International University
alokdasshuvo46@gmail.com

Swakkhar Shatabda
United International University
swakkhar@cse.uiu.ac.bd

***Abstract*—Financial market participants continuously watch financial and economic news headline to ectract relative information.The efficient market theory states that stock prices reflect all historical knowledge, and that new information is instantly included when establishing future stock values. Therefore, quick extraction of positive or negative sentiments from news is crucial for traders, portfolio managers, and investors making investment decisions.However, compared to the English language, only a few study has been conducted in the Bangla language field due to data unavailability and language complexity. The study aims to identify investor sentiment expressed to Bengali-language financial news Headlines. In our study, we use Bangla newspaper Headlines and generate predictive classification models. Support Vector Machine (SVM), Logistic Regression and Convolutional Neural Network(CNN) a used for Classification.We have gained highest F1-Score 78.0 % from SVM model. As Bengali-speaking financial markets become more active and prominent, the outcome of this study might be helpful to investors, financial institutions, and governments.**

***Index Terms*—Sentiment analysis, Financial Sentiment, Bangla NLP, Tf-Idf, N-grams, SVM, CNN;**

I. Introduction

Sentiment analysis also known as opinion mining, is a natural language processing technique that categorizes subjective feelings, opinions, and attitudes in text, such as documents, social media posts, or news articles, as positive, negative, neutral [11]. Financial sentiment analysis is a specialized field that uses natural language processing and machine learning to analyze the emotional tone in financial data, news, and reports, determining positive, negative, or neutral sentiment [3].With its data-driven approach to assessing investor and market mood, financial sentiment analysis is essential to the financial sector [16]. Better decision-making, risk management, and market adaptation are all facilitated by it. The value of financial sentiment analysis is expected to increase more as technology and data analysis methodologies develop.

The Manual analysis of financial sentiment has several limitations, including subjectivity, scalability time sensitivity, cost, human error and emotional fatigue [7]. The analysis may not cover a wide range of data sources, leading to incomplete insights. Additionally, the analysis may not provide real-time insights, potentially causing delays in decision-making. Furthermore, the analysis may not cover historical data, making it challenging to perform historical trend analysis or back-test trading strategies. Furthermore, the lack of standardization and limited data volume can further complicate the process.

Machine learning is a powerful tool for sentiment analysis in financial markets [20]. It offers scalability, consistency, real-time analysis, pattern recognition, multilingual support, sentiment classification and back-testing. These algorithms process vast amounts of textual data from various sources, enabling investors and traders to react quickly to market-moving news and events. They can also identify complex patterns and relationships in data, revealing correlations between sentiment and market movements [21], [22], [24]. Furthermore, machine learning models can be customized to focus on specific financial domains or tailor sentiment analysis to unique requirements. But Sentiment analysis in the Bengali language faces challenges due to its unique characteristics, data availability, resource constraints, market influence, research priorities, and language resources [5]. Understanding Bengali linguistic intricacies is crucial for analyzing sentiment, while obtaining comprehensive sentiment-labeled datasets can be challenging. Additionally, developing sentiment analysis tools requires significant resources, making Bengali a less studied language.

The goal of this study was to extract the sentiment related to financial news headlines an build a good classification model using Bangali financial dataset. For that purpose, we have

- collected financial related news headlines from reliable and well-known Bengali newspaper.
- analyze and tried to find out important feature regarding the dataset
- used Support Vector Machine (SVM), Logistic Regression and Convolutional Neural Network(CNN) for building classification model
- made prediction model which can determine whether a Bengali news headline has got positive or negative sentiment
- given brief details about our approach which might help future researcher

II. Related Work

Sentiment analysis in Bangla language is underresearched due to lack of a large and standard dataset, resulting in insufficient findings for future research compared to other languages. The only existing work done on Bangla financial sentiment analysis is by Kibtia *et al.* [4]. The dataset they used is collected from Kaggle and ”Somoi” television channels. To demonstrate the efficiency of the model, they applied various machine-learning techniques with vector-based feature encoding. There are 615 data points consisting of 350 positive and 265 negative data points. Word2vec and word embedding are used to preprocess the dataset. With an accuracy rate of 83%, they concluded that the Random Forest Algorithm (RFA) is the most efficient technique.

Dogu Tan Araci introduced FinBERT, a language model based on BERT, to tackle NLP tasks in the financial domain [2]. They used two datasets: TRC2-financial which is collected from the Reuters website and also Financial PhraseBank Dataset which is from LexisNexis database. They build 7 models, including LSTM, LSTM with ELMo, FinBERT, etc. FinBERT outperforms other models with an accuracy of 86% in all data. Bert tokenization technique has been used for text representation.

Cheng Qian *et al.* applied deep neural network to analyze financial sentiment from the dataset collected from twitter [18]. A total of 290,282 English tweets were collected for this purpose. The globe model was used for text representation and they achieved an analysis accuracy of 89.21%.

Frank *et al.* try to find out the error patterns of some well-known sentiment analysis techniques rather than presenting a new model [26]. They used 8 models including SVM, bi-LSTM, S-LSTM, BERT etc. to find out the accuracy and error patterns of the models.They achieved the highest 95.7% accuracy on the Yelp dataset and 76.9 % accuracy on the Stocksen dataset using the BERT model.

On the microblogs and headlines of the financial domain, Hitkul *et al.* conduct aspect-based sentiment analysis [13]. FiQA 2018 dataset was used in this study. A multi-channel convolutional nueral network was used to analyze the sentiment and bi-LSTM model was used to astract aspects. They achieved f1-score of 69% by using bi-LSTM model.

Akhtar *et al.*offer a novel method using a Multi-Layer Perceptron (MLP) network to analysis the financial sentiment [1]. CNN, GRU and LSTM were used as deep learning models. These deep learning models were combined with a traditional supervised model based on Support Vector Regression (SVR) to create an ensemble model. They used two types of datasets: Microblogs and News headlines. Models were evaluated using the SemEval-2017 dataset. They gained an accuracy of 79.7% on microblogs and 78.6% on news dataset repectively.

Ling Luo *et al.* introduce a model named FISHQA for financial sentiment analysis [17]. It uses a hierarchical structure with attention mechanisms from several granularities, including both word level and sentence level. They collected about 30, 000 documents as dataset from different Chinese mainstream. Their model FISHQA-Q1 surpasses the rest of the model with an accuracy of 94.46%.

In our study, we present a fresh set of 1836 financial-related news headlines that have been carefully processed, properly verified by multiple sources, and prepared for tests and sentiment analysis implementation. We analyzed and tried to find important findings from the dataset. We implement three machine learning approach to build classification model for sentiment prediction. We provide comprehensive information about our methodology, experiments, and result analysis in order for other researchers to learn from it in the future.

III. Materials and Methods

A. Dataset Details

The dataset we used for the study is collected manually from Prothom Alo newspaper as it is one of the reliable resource for bangla news. Only news Headlines related to financial sentiment are collected within the date range of 01 January 2020 to 20 August 2023. This dataset is merged with the dataset used by Kibtia *et al.* to increase the dataset instances [4]. There is total number of 1836 news headlines which consists positive 888 headlines denoted by 1 and negative 948 headlines denoted by 0. Figure 1 represents the data distribution of our collected corpus and Figure 2 presents the sample of our dataset .

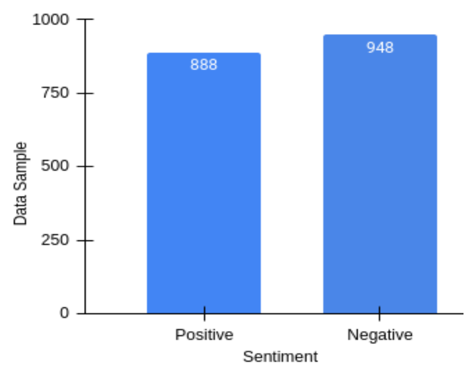

Fig. 1. Data distribution in our dataset

Headline	label
এক দশকের মধ্যে সর্বনিম্ন প্রবৃদ্ধি	0
রিজার্ভ কমে আবার ৩০ বিলিয়নের নিচে	0
৩৩ বছরে দ্বিতীয় সর্বোচ্চ বিদেশি বিনিয়োগ	1
সংকটেও রেকর্ড মুনাফা লেনদেনে ফিরল শেয়ার	1
সবচেয়ে বড় দুই বাজারে পোশাক রপ্তানি কমেছে	0
রপ্তানি আয়ে সুখবর প্রবৃদ্ধি ৬.৬৭%	1

Fig. 2. Sample Data

B. Methodology

We collected financial news headlines from Prothom Alo newspaper and lebelled it manually. After prreprocessing the dataset, we apply tf-idf vectorization for both SVM and Logistic Regression and word-embedding technique for CNN. We then split the dataset with a ratio of 80:20 for train and test dataset. Train dataset is used to train the model and test dataset is used to evaluate the model.Then finally build prediction model which can classify from any bangali news headlines wheather it belongs to positive sentiment or negative. Figure 3 represents the methodology of our study.

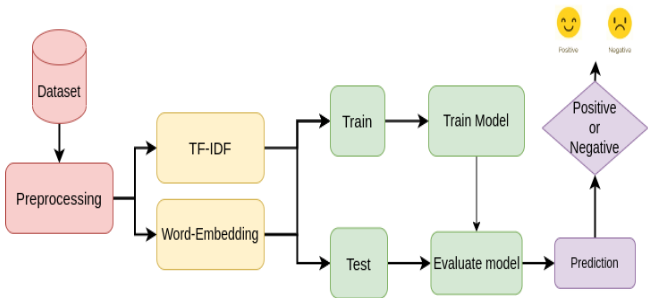

Fig. 3. Methodology

C. Pre-processing

NLP involves pre-processing raw text data for improved data quality, reduced noise, and consistent representation for tasks like text classification, sentiment analysis, and machine translation [15]. In pre-processing steps we cleaned our dataset by removing duplicate value,punctuation marks etc. For both SVM and Logistic Regression model, out corpus is needed to transform into tf-idf vectorization [6]. Unigram technique is applied to convert our corpus into tf-idf vectorization.Similarly word-embedded vectorization technique is applied to our corpus for applying CNN [14].

D. Proposed Model and Training

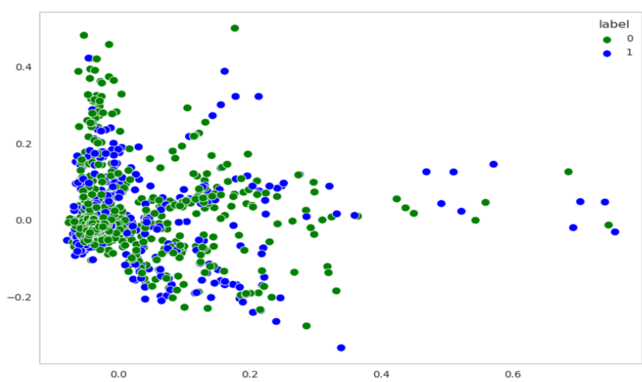

Fig. 4. Scatter-Plot of our Dataset

We plot a Scatter-plot of our Data which is shown in Figure 4. From the Plot we can clearly understand that there can be no linear line which can divide both class separately. This is an Ideal situation for apply non-linear method like CNN or SVM with Non-linear configurations. So for our study, we selected both SVM and CNN as non-linear method and Logistic Regression as linear method [19], [25].

- Support Vector Machine (SVM) and Logistic Regression (lr): Before fit in the models, the data imbalance is checked. SMOTE technique is applied to balance the train dataset [23]. After pre-processing steps, the tf-idf feature vector size was 3622. Principal Component Analysis (PCA) technique is applied and feature size is reduced to 1500 [9]. GridSearchCV technique is applied to find the best hyperparameters for the model [10]. Final model is defined using those best hyperparameters.
- Convolutional Neural Network (CNN): A model with 1,847,286 trainable parameters and 1,856 Non-trainable parameters is introduce to deal with the dataset. Batch normalization and Dropout is used in every convolutional layer [8]. The ouput of convolutional layer i flatterned before entering into Dense layer. Relu is used as activation function in every layer expect classification layer. Sigmoid used in final classification layer as activation function. Adam is used as optimizer with initial learning rate 0.001. Categorical Crosentrophy is used as loss fucntion. Batch size is used 16 while number of epoch is selected as 100. A custom callback function is introduced containing reducing learning rate, early stopping and checkpoint function [12]. The architecture of the CNN is shown in Figure 5.

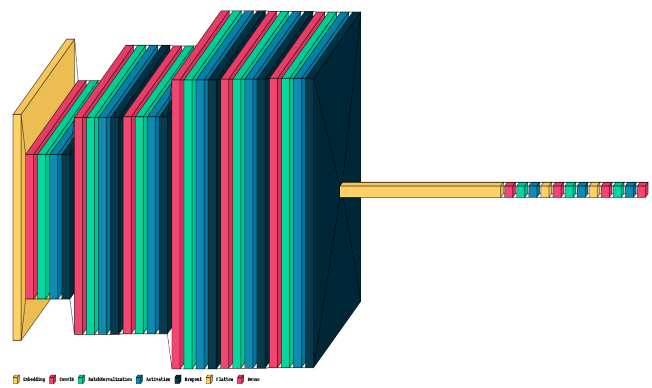

Fig. 5. Architecture of CNN

The Dataset is divided into two set: train set (80%) and test set (20%). The model is trained on train set and evaluated on test dataset.

IV. Result Analysis

Our models are evaluated using accuracy and f1-score matrices. Equation 1,2,3,4 represents accuracy,precision,recall and f1-score equation where TP indicates true positive, TN indicates true negatve, FN inidcates false negative and FP indicates False Positive value predicted by the models.

Model accuracy is a performance metric in machine learning and classification tasks, indicating how many correctly predicted instances out of a dataset match the actual outcomes.

$$Accuracy = TP + TN/(TP + FP + TN + FN) - -(1)$$

Precision is crucial for high false positives or accurate predictions, often used in combination with recall and F1 score for comprehensive evaluation of model performance in imbalanced datasets.

$$precision = TP/(TP+FP) - - - - - - - - - - - - - -(2)$$

Recall is crucial in avoiding false negatives and high-cost cases like medical diagnoses and anomaly detection. It's often used alongside precision and F1 score to evaluate a model's performance in imbalanced datasets.

$$Recall = TP/(TP + FN) - - - - - - - - - - - - - - (3)$$

The F1 score is a metric that balances precision and recall in a model, ensuring accurate predictions while avoiding false positives and negatives.The F1 score can be a more informative metric than accuracy for evaluating a model's performance in situations where false positives and false negatives have different consequences.

$$F1Score = 2((Precision*Recall)/(Precision+Recall))-(4)$$

Our SVM model acquire highest 78% f1-score while logistic regression model and CNN achieve 75.7% and 67.8% respectively.Our Logistic Regression model acquired highest accuracy of 76.6% compared to 76.3%, and 67.9% from SVM and CNN model. The comparison between f1-score is shown in Figure and between accuracy is shown in Figure .

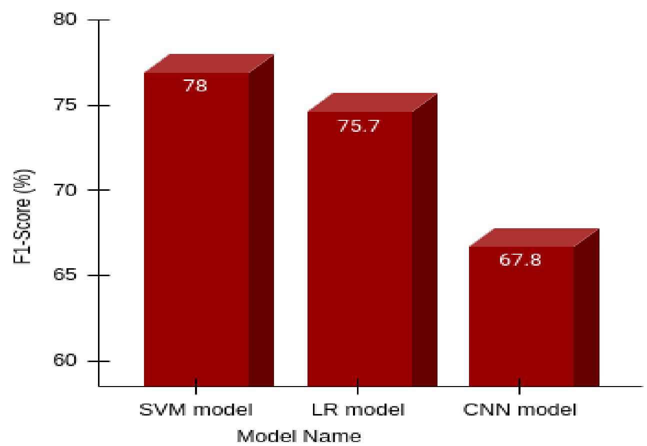

Fig. 6. F1-Score comparison between models

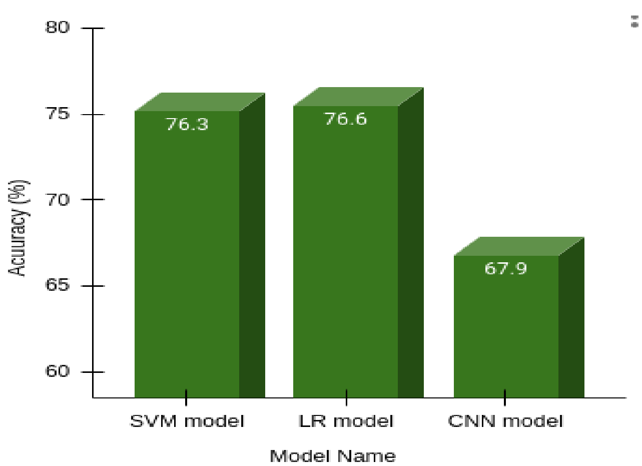

Fig. 7. Accuracy comparison between models

SVM model performs better compared to others models in Both f1-score and accuracy. Linear Regression model also does well on prediciting sentiment correctly. But CNN model has got lowest accuracy and f1-score. This could be because of less number of dataset compared to the parameter of CNN. CNN shows tendency of memorizing the data when dataset is much lower than the training parameter. A more simple model or with more training dataset may increase the ability of CNN to classify financial sentiment more accurately.

V. Conclusion

Financial sentiment analysis is crucial for stakeholders in the industry, enabling informed decisions, risk management, and market opportunities identification. It enhances strategies, monitors market activity, and helps companies assess brand perception, engage with shareholders, and tailor reporting to audience interests. In this study our goal was to build a classification model for bangla financial sentiment. We review the recent work related to this topic and explored relative issues behind less study done in bangla financial sentiment. We succesfully build a classification model with good accuracy and f1-score. we also provide detailed description of our study. In future, we will try to explore more using LSTM,Transfer Learning with larger Dataset related to Bangla financial Sentiment.

For Project:

Visualization: https://www.kaggle.com/code/shuvoalok/visualization-fsa/notebook

SVM and LR model: https://www.kaggle.com/code/shuvoalok/fsa-ml-with-pca/notebook

CNN model: https://www.kaggle.com/code/shuvoalok/fsa-nn/notebook

References

[1] Md Shad Akhtar, Abhishek Kumar, Deepanway Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 540–546, 2017.

[2] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[3] Sven-Kristjan Bormann. Sentiment indices on financial markets: What do they measure? Technical report, Economics Discussion Papers, 2013.

[4] Kibtia Chowdhury and Swakkhar Shatabda. Sentiment analysis on bangla financial news. In *2021 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 64–67. IEEE, 2021.

[5] Shaika Chowdhury and Wasifa Chowdhury. Performing sentiment analysis in bangla microblog posts. In *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, pages 1–6. IEEE, 2014.

[6] Mamata Das, PJA Alphonse, et al. A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset. *arXiv preprint arXiv:2308.04037*, 2023.

[7] Chedia Dhaoui, Cynthia M Webster, and Lay Peng Tan. Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing*, 34(6):480–488, 2017.

[8] P Dileep, Dibyaiyoti Das, and Prabin Kumar Bora. Dense layer dropout based cnn architecture for automatic modulation classification. In *2020 national conference on communications (NCC)*, pages 1–5. IEEE, 2020.

[9] Reza Drikvandi and Olamide Lawal. Sparse principal component analysis for natural language processing. *Annals of data science*, 10(1):25–41, 2023.

[10] Mohd Anul Haq, Mohd Abdul Rahim Khan, and Mohammed Alshehri. Insider threat detection based on nlp word embedding and machine learning. *Intell. Autom. Soft Comput*, 33:619–635, 2022.

[11] Eduard H Hovy. What are sentiment, affect, and emotion? applying the methodology of michael zock to sentiment analysis. *Language production, cognition, and the Lexicon*, pages 13–24, 2015.

[12] Asmida Ismail, Siti Anom Ahmad, Azura Che Soh, Khair Hassan, and Hazreen Haizi Harith. Improving convolutional neural network (cnn) architecture (minivggnet) with batch normalization and learning rate decay factor for image classification. *International Journal of Integrated Engineering*, 11(4), 2019.

[13] Hitkul Jangid, Shivangi Singhal, Rajiv Ratn Shah, and Roger Zimmermann. Aspect-based financial sentiment analysis using deep learning. In *Companion Proceedings of the The Web Conference 2018*, pages 1961–1966, 2018.

[14] Qilu Jiao and Shunyao Zhang. A brief survey of word embedding and its recent development. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, pages 1697–1701. IEEE, 2021.

[15] Subbu Kannan, Vairaprakash Gurusamy, S Vijayarani, J Ilamathi, Ms Nithya, S Kannan, and V Gurusamy. Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2014.

[16] Srikumar Krishnamoorthy. Sentiment analysis of financial news articles using performance indicators. *Knowledge and Information Systems*, 56(2):373–394, 2018.

[17] Ling Luo, Xiang Ao, Feiyang Pan, Jin Wang, Tong Zhao, Ningzi Yu, and Qing He. Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention. In *IJCAI*, pages 4244–4250, 2018.

[18] Cheng Qian, Nitya Mathur, Nor Hidayati Zakaria, Rameshwar Arora, Vedika Gupta, and Mazlan Ali. Understanding public opinions on social media for financial sentiment analysis using ai-based techniques. *Information Processing & Management*, 59(6):103098, 2022.

[19] Chaudhary Jashubhai Rameshbhai and Joy Paulose. Opinion mining on newspaper headlines using svm and nlp. *International journal of electrical and computer engineering (IJECE)*, 9(3):2152–2163, 2019.

[20] Thomas Renault. Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digital Finance*, 2(1-2):1–13, 2020.

[21] Nusrat Rouf, Majid Bashir Malik, Tasleem Arif, Sparsh Sharma, Saurabh Singh, Satyabrata Aich, and Hee-Cheol Kim. Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions. *Electronics*, 10(21):2717, 2021.

[22] Shunrong Shen, Haomiao Jiang, and Tongda Zhang. Stock market forecasting using machine learning algorithms. *Department of Electrical Engineering, Stanford University, Stanford, CA*, pages 1–5, 2012.

[23] Sahar Tahvili, Leo Hatvani, Enislay Ramentol, Rita Pimentel, Wasif Afzal, and Francisco Herrera. A novel methodology to classify test cases using natural language processing and imbalanced learning. *Engineering applications of artificial intelligence*, 95:103878, 2020.

[24] Franco Valencia, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy*, 21(6):589, 2019.

[25] Bhartendoo Vimal and S Anupama Kumar. Application of logistic regression in natural language processing. *Int J Eng Res*, 9, 2020.

[26] Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. Financial sentiment analysis: an investigation into common mistakes and silver bullets. In *Proceedings of the 28th international conference on computational linguistics*, pages 978–987, 2020.