ISSN 2347 - 3983

Volume 11. No.5, May 2023

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter021152023.pdf

https://doi.org/10.30534/ijeter/2023/021152023

A Voting classification approach for Sentiment Extraction from Bengali text

Piyal Roy[1], Amitava Podder[2*], Smaranika Roy[3]

[1]Department of Computer Science & Engineering, Brainware University, India, piyalroy00@gmail.com

[2*]Department of Computer Science & Engineering, Brainware University, India, amitavapodder24@gmail.com

[3]Department of Computer Science, Sarada Ma Girls' College, India, smaranika0045@gmail.com

Received Date: March 27, 2023 Accepted Date: April 17, 2023 Published Date : May 07, 2023

ABSTRACT

Sentiment extraction is one of the most challenging tasks in Natural Language Processing (NLP). It is essential for analysing consumer and user feedback on social media sites and in the commercial world. Finding sentiments or emotions in raw text data and identifying their polarity, or whether they are positive or negative, is the main objective of sentiment extraction. This area has been the focus of various research projects for English and other significant natural languages. In this article, we offer a voting classification method that uses a variety of machine learning classifiers to extract sentiment from Bengali language text. We explored Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier, Multinomial Nave Base and Ridge Classifier, and lastly, we used a voting classification strategy to extract sentiments from social media comments.

Key words : Sntiment Extraction, Logistic Regression, Support Vector Classifier, Multinomial Naïve Base, Decision Tree, Voting Classification

1. INTRODUCTION

A. *Context*

Over the past decade, researchers have relentlessly conducted extensive research on sentiment extraction or sentiment categorization. It is the process[6] by which, using various text classification approaches, we may extract and categorize feelings from a particular document, paragraph, phrase, or clause. The technique of extracting opinions, sentiments, attitudes, emotions, etc. from textual data and categorizing them according to their polarities is known as sentiment extraction, also known as sentiment analysis or opinion mining [9][13]. In a different research [7], the authors state that the process of sentiment analysis entails determining the predominant emotion in a given text, which is often characterized as a categorical variable with three possible values: positive, negative, and neutral. The text data for the sentiment analysis task can be gathered from a [14] variety of sources, including social media (e.g., Facebook, Twitter, Instagram, LinkedIn), various product review websites (e.g., Amazon), blogs, news articles and comments on them, YouTube comments, movie review websites (e.g., Yelp, IMDB, Rotten Tomatoes), online forum discussions, emails, customer satisfaction surveys, and more.

B. *Importance*

Today, it has become clear that it is important for both individuals and major business and political organizations to comprehend the emotions that people exhibit in a variety of scenarios. Customers now have more alternatives than ever for expressing their thoughts and feelings openly and extensively. Comments may be made practically anywhere, including movie reviews, online stores, software archives, and social networking platforms. Researchers are offering a variety of methods for analyzing these comments in order to ascertain their emotions and the appropriate interpretation. Sentiment extraction is used to systematically examine the market strategy based on consumer input. However, when examining how the general population feels about government initiatives, sentiment extraction is crucial. Additionally, a variety of sentiment extraction tools let users find the right scripts and validate them in the code repository using code reviews, find unfriendly online communities, and do much more.

C. *Contribution*

This paper presents an approach for sentiment analysis for Bengali text data using a voting classification technique. Our approach is applied on a benchmark dataset which can be downloaded from internet[1] and obtained an accuracy of 81%.

[1] https://www.kaggle.com/datasets/cryptexcode/sentnob-sentiment-analysis-in-noisy-bangla-texts?resource=download

D. Organisation of the paper

Section 2 discusses some related work on sentiment extraction on texts in various languages. Section 3 presents some challenges of sentiment extraction. Our approach for this paper is introduced in section 4. Next, Section 5 depicts the dataset used in our experiment. Then we present the experimental outcomes in Section 6. Finally, Section 7 draws our conclusions and outlines for the future research.

2. RELATED WORK

The Bengali language has a rich history and a significant cultural effect. There are 228 million native speakers and 37 million people learning it as a second language globally. It is still challenging to do research on sentiment extraction for Bengali text data since there aren't any suitable NLP tools (such stemmers, pos-taggers, lemmatizers) for the Bengali language. The majority of study on sentiment extraction has been done for the English language, however there has been relatively little done for Bengali.

A method for classifying sentiment in Bengali tweets was given by Kamal Sarkar [8] utilising deep neural networks, which include three layers: a convolutional layer, a hidden layer, and a softmax layer as the output layer. In another study, Attia et al. [2] suggested a language-independent model for multi-class sentiment analysis utilising a simple neural network design with five layers (Embedding, Conv1D, GlobalMaxPooling and two Fully-Connected). Their suggested model does not rely on linguistic characteristics such as ontologies, dictionaries, or morphological or syntactic pre-processing. Aziz et al. [10] introduced a novel approach for classifying Bengali text's multi-modal sentiment using RNN. They were 85.67% accurate overall. Another multilingual technique for the sentiment analysis challenge was described by Sazzed et al [9]. They applied a variety of machine learning techniques on a Bengali corpus and its automatically translated English equivalent, including Logistic Regression (LR), Ridge Regression (RR), Support Vector Machine (SVM), Random Forest (RF), Extra Randomized Trees (ET), and LSTM. Other models we proposed by Phani et al. [7] and Vilares et al. to perform multilingual sentiment analysis using twitter data. In order to do sentiment analysis on twitter data, Nandan et al. [12] presented a hybrid strategy employing both lexicon-based and machine learning approaches. For the multilingual identification [11] of hate speech against women and immigrants in twitter data, a method was put out by Basile et al. [3].

Borele et al. [4] identified the distinguishing characteristics of existing solutions in their study of several machine learning-based systems for sentiment analysis operations. Habimana et al. [5] performed a detailed study of deep learning techniques that have been used for various sentiment analysis applications. Additionally, their survey offers a specific dataset at the conclusion of each sentiment analysis job. The review identifies the problems that are currently plaguing society and suggests potential remedies.

3. CHALLENGES

Data on the relevant topic is required for NLP research, and machine learning need a lot of data for training the models. NLP models improve in accuracy when more data is used to train them. This criterion is not satisfied in the case of a low resource language like Bengali. In order to train the model for sentiment extraction, a large volume of reviews or comments are required, however for the Bengali language, the researchers encounter the following challenges.

- Incomplete Sentence: People mostly give their reviews in incomplete sentences which sometimes makes the reviews meaningless.
  e.g. আমি একজন সাধারন নাগিরক। আয়কর জমা তো দূরের কথা তার ধারে কাছে

- Miss-spelled words: The reviews given on the online platforms are full of spelling mistakes. These misspelled words sometimes make it hard to understand the meaning of the review. e.g. ইসমার্ট ফনের ব্যবহার আমাদের শরীর র ক্ষতি করছে।

- Emojis and Punctuation Symbols: Nowadays, using various punctuation marks and emojis in place of actual words has grown popular. Although these non-characters are understandable to humans, when it comes to training an NLP model, they have no significance whatsoever.
  e.g. শুভ জন্মিদন.:):):)

4. METHODOLOGY

Several deep learning models have been used for this study on sentiment extraction of Bengali texts. The deployed models are Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier, Multinomial Naïve Base and Ridge Classifier. Our approach is divided into five different segments.

A. Data Preprocessing

First, we first purge all empty comments, stop-words and unnecessary punctuation marks, leaving only the meaningful words. It guarantees the model's potential for performance improvement and increases classification accuracy. Stop words and punctuation are eliminated by well-known search engines like Google and Yahoo to enable quick and accurate data retrieval from their database.

Our observations lead us to the conclusion that some reviews are both too short and too long. In order to address this issue, we merely pad or truncate all of our reviews to a predetermined fixed length. Reviews that are shorter than the predefined length have 0s appended to them, while longer reviews have their first 100 words kept and the remaining words are trimmed.

B. *Data Vectorization*

We vectorize the text data by converting each string of text into a series of integers using TF-IDF[2] vectorizer. Furthermore, vectorization will speed up calculation compared to non-vectorized implementation. The comment labels are converted into numbers as follow-

- Positive = 1
- Negative = 2

C. *Split Dataset*

At the next phase, we split the dataset into two subsets. We call the first subset as train data and the second subset as test data. Train data contains 90% of the total dataset chosen randomly. Test data contains the remaining 10% of the actual dataset and is used for testing the predictions made by the model.

D. *Experimental Setup*

All the evaluated models are implemented using Python-3 language and executed on the Google Colab Platform with NVIDIA-SMI GPU and 32 GB of RAM. We have used Scikit-learn[3] which is a free machine learning library for Python programming. It features various classification, regression and clustering algorithms.

E. *Training & Testing*

We randomly picked 90% of the dataset to train the models. After successful training, we used the remaining 10% of the dataset to test the trained models.

5. DATASET

It is really difficult to find benchmark dataset for Bengali sentiment analysis work. However, most researchers do not make their datasets publicly accessible along with their research outcomes. As a result, some researchers choose building their own dataset from scratch, which is time-consuming. We conducted extensive web searches and came upon a benchmark dataset provided by Khondoker et al [1]. This dataset, which was gathered from several social media platforms, is a compilation of unofficial public comments on news articles and videos. The collection consists of 13 distinct areas, including politics, national security, sports, food, international trade, technology, entertainment, business, lifestyle, education, travel, fashion, and agriculture. Three separate annotators are used to assign one of the three sentiment labels—positive, negative, or neutral—to each occurrence of the dataset. The dataset was prepared using the necessary annotation criteria by ten undergraduate students. A majority vote is used to determine the final label for each comment in the dataset. A detailed statistical analysis of the dataset is given in Table I.

Table 1: Dataset Statistics

Label	Comments	Average Words	Average Sentences
Positive	5,709	1.64	16.33
Negative	6,410	1.73	15.88
Neutral	3,609	1.45	12.94
Overall	15,728	1.63	15.37

A sample of the dataset can be found in Table 2.

Table 2: Sample Dataset

Comments	Labels
আপনার নাচ দেইখা আমারো খাইতে ইচ্ছা করছে	*Positive*
যেমন : পরীক্ষার রেজাল্টের সময় , বিভিন্ন ব্যানিজ্যিক প্রচার ইত্যাদি	*Neutral*
মারা খেয়ে গেছি । এই খাবার বাঙ্গালী জীভের জন্য উপযুক্ত নয় মোটেই	*Negative*
ভাই এটা কি পুরা সাপ্তাহেই পাওয়া যায় নাকি কোন দিন অফ থাকে	*Positive*
এই ধরনের খাবার ভিডিও রাতে শুয়ে শুয়ে দেখলে কেমন লাগে ভাবতে পারেন	*Positive*
আমার মত কে কে কমেন্ট পড়তে আসছ হাত তুলো । ✋	*Neutral*
তোমরা যদি শব্দ উচ্চারণ না করতে জান তাহলে খবর করবে না তার কারণ হল সে হল গলি বয় ।	*Negative*
আসলে রাস্তাগুলো দেখে আমার ভয় লাগছিল। দোয়া রইলো আপনাদের পতি	*Positive*
এখানে খেলে মনে হয় না যে কোন হোটেলে খেয়েছি। এটাই বিউটি বোর্ডিংয়ের প্লাস পয়েন্ট	*Positive*

6. RESULT

Our approach uses four machine learning models i.e., Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier, Multinomial Naïve Base and Ridge Classifier. We have evaluated the performance of our approach by using Precision, Recall F-Score and Accuracy metrics. The detailed evaluation result is given in Table 3.

The metrics have been calculated using the following formulas –

[2] http://www.tfidf.com/

[3] https://scikit-learn.org/stable/

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Where,

TP = True Positive Samples
TN = True Negative Samples
FP = False Positive Samples
FN = False Negative Samples

Table 3: Result Analysis

Model	**Precision**	**Recall**	**F-Score**	**Accuracy (%)**
Logistic Regression	0.78	0.79	0.78	79
Decision Tree Classifier	0.77	0.77	0.77	77
Random Forest Classifier	0.76	0.77	0.77	76
Support Vector Classifier	0.76	0.75	0.75	75
Multinomial Naïve Base	0.71	0.71	0.71	71
Ridge Classifier	0.70	0.69	0.70	70
Voting Classifier	**0.81**	**0.80**	**0.81**	**81**

Therefore, it can be seen that, our approach outperforms all the individual models and achieves an accuracy of 81%.

7. CONCLUSION

In this paper, we have presented a voting classification approach for extracting sentiments from Bengali comments in different social media sites. The absence of benchmark datasets for sentiment extraction in Bengali is one of the major issues we encountered while doing our research. The used dataset contains a lot of noise, which reduces the overall accuracy. In order to develop a more precise sentiment extraction model for Bengali text, we aim to rectify the deficiencies in the training data and build our own dataset.

REFERENCES

[1] Islam, K. I., Kar, S., Islam, M. S., & Amin, M. R. (2021, November). SentNoB: A Dataset for Analysing Sentiment on Noisy Bangla Texts. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 3265-3271).

[2] Attia, M., Samih, Y., Elkahky, A., & Kallmeyer, L. (2018, May). Multilingual multi-class sentiment classification using convolutional neural networks. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

[3] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., ... & Sanguinetti, M. (2019, June). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th international workshop on semantic evaluation (pp. 54-63).

[4] Borele, P., & Borikar, D. A. (2016). An approach to sentiment analysis using artificial neural network with comparative analysis of different techniques. IOSR Journal of Computer Engineering (IOSR-JCE), 18(2), 64-69.

[5] Habimana, O., Li, Y., Li, R., Gu, X., & Yu, G. (2020). Sentiment analysis using deep learning approaches: an overview. Science China Information Sciences, 63(1), 1-36.

[6] Liao, C. W., Wang, I. C., Lin, K. P., & Lin, Y. J. (2021). A fuzzy seasonal long short-term memory network for wind power forecasting. Mathematics, 9(11), 1178.

[7] Phani, S., Lahiri, S., & Biswas, A. (2016, December). Sentiment analysis of tweets in three Indian languages. In Proceedings of the 6th workshop on south and southeast asian natural language processing (WSSANLP2016) (pp. 93-102).

[8] Sarkar, K. (2019). Sentiment polarity detection in Bengali tweets using deep convolutional neural networks. Journal of Intelligent Systems, 28(3), 377-386.

[9] Sazzed, S., & Jayarathna, S. (2019, July). A sentiment classification in bengali and machine translated english corpus. In 2019 IEEE 20th international conference on information reuse and integration for data science (IRI) (pp. 107-114). IEEE.

[10] Sharfuddin, A. A., Tihami, M. N., & Islam, M. S. (2018, September). A deep recurrent neural network with bilstm model for sentiment classification. In 2018 International conference on Bangla speech and language processing (ICBSLP) (pp. 1-4). IEEE.

[11] Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2017). Supervised sentiment analysis in multilingual environments. Information Processing & Management, 53(3), 595-607.

[12] Nandan, M., Chatterjee, S., Parai, A., & Bagchi, O. (2022). Sentiment Analysis of Twitter Classification by Applying Hybrid-Based Techniques. In Proceedings of the 3rd International Conference on Communication, Devices and Computing (pp. 591-606). Springer, Singapore.

[13] Lakshmikantham, V., & Devi, J. V. (2008). Theory of fractional differential equations in a Banach space. *European Journal of Pure and Applied Mathematics*, 1(1), 38-45.

[14] Podder, A., Roy, P., & Roy, S. (11 2022). Steganography Techniques - An Overview. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 323–327. doi:10.32628/CSEIT228642