Sentiment Analysis of Banglish Food Reviews using Natural Language Processing Techniques

No Author Given

No Institute Given

Abstract. Food review analysis can play an important role in the restaurant business, where business owners can improve their service through the sentiments of the reviews. In Bangladesh, customers often share their reviews on social media or restaurant websites. Most of the time, they use Banglish sentences because they are easy to write. Banglish is Bangla written in English words or phrases. This project aims to perform sentiment analysis on food reviews written in Banglish, a hybrid language that combines elements of both Bengali and English. The project will utilize Natural Language Processing (NLP) techniques to analyze the reviews and classify them into positive, negative, or neutral sentiment categories. So we gathered Banglish food reviews from social media and different websites. Selenium, Beautiful Soup, and the instant data scrapper tool for collecting data. Our model To train our dataset, we used BERT, Robert, LSTM, and Naive Bayes. We evaluated the performance of each model and found that the Naive Bayes model had the highest accuracy of 86%. The BERT model had an accuracy of 85%, while the LSTM and Roberta models had accuracies of 73% and 74%, respectively. These models can be used by businesses in the food industry to better understand their customers' feedback and improve their products and services.
Keywords: Natural Language Processing (NLP); AI safety; Sentiment Analysis; Banglish food reviews

1 Introduction

A method of Natural Language Processing (NLP) called sentiment analysis recognizes the emotional undertone of a document. It is a well-liked method for gathering feedback on a concept or item for any company. Numerous variables, including artificial intelligence and machine learning, are used to scan text for sentiment and determine the level of sentiment to determine the types of sentiments it is conveying.

The most impactful side of sentiment analysis is that it helps collect insights into real-time customer sentiment and customer experience. To evaluate online sources like emails, blog posts, online tickets, news stories, online communities, comments, etc., these tools typically use text analytics. Also, algorithms are used to determine whether the customer is expressing positive or negative words by implementing rule-based hybrid methods of scoring.

A sentiment analysis follows a bunch of steps to analyze a piece of text using a machine-learning model for human language. First of all, it collects data by identification, and then it cleans that data to remove noise and irrelevant content. The next step is extracting features, which uses a bag of techniques to extract text features for the identification of negative or positive sentiments. After that, it picked an appropriate ML model and then classified the sentiment. Sentiment analysis systems fall into multiple categories, such as fine-grained sentiment analysis, emotion detection analysis, intent-based analysis, and aspect-based analysis.

An organization must primarily comprehend the needs and desires of its consumers in order to effectively market its goods. Organizations do not simply review every single bit of info. They, therefore, employ sentiment analysis to examine client reviews found in online sources. Additionally, it tracks the state of the industry and assesses the effectiveness of marketing initiatives. Although there are many advantages for organizations, there are still some difficulties, including neutral feelings, ambiguous language, unclassifiable language, ambiguous sentiments, named-entity identification, limited data sets, language development, fake evaluations, and the need for human involvement. Despite its difficulties, sentiment analysis has enormous potential for some industries.

The major purpose of this research is to learn how to use sentiment analysis in microblogging to examine user reactions to a company's wares. This research shows how an NLP and Bidirectional LSTM model can classify the sentiments of the tweets of users optimally.

2 Related Works

As a consequence of its enormous influence on the competitive business sector, product market demand analysis is important for developing company strategies [1]. According to their discussions, the importance of product market demand analysis for business strategies is particularly apparent in the context of the Bangladeshi market for smartphones. The majority of the population speaks Bengali and uses Banglish text to interact on social media, making it a critical source of data for assessing market demand. They collected data from social media platforms and other websites and used natural language processing (NLP) techniques, such as sentiment analysis and named entity identification, to analyze the data. They trained their datasets with machine learning models, such as Spacey's custom NER model and Amazon Comprehend Custom NER, and deployed a Tensorflow sequential model with parameter tweaking for sentiment analysis. The model had an accuracy of 87.99% in Spacy Custom Named Entity Recognition, 95.51% in Amazon Comprehend Custom NER, and 87.02% in the sequential model for demand analysis.

The authors aimed to identify the most popular smartphones by gender and provide entrepreneurs with a statistical and realistic market demand analysis. There are challenges in collecting and labeling the Banglish text data set and the importance of sentiment analysis and natural language processing for obtaining, quantifying, and analyzing consumer preferences. ML advancements throughout the previous decade, particularly in recent years, have resulted in the dominance of ML-based approaches for the majority of NLP tasks. The findings of an assessment study of lexicon-based sentiment analysis resources for German texts are presented [2]. After analyzing 20 sentiment-annotated corpora and 19 sentiment lexicon resources for German texts from various topics, they focus on identifying and analyzing human sentiment and emotions across a range of application domains [2]. It emphasizes text as a modality and categorizes texts of varying lengths according to the polarity represented in the text, which refers to whether a text's attitude is more strongly positive or negative. According to the author, part-of-speech (POS) information can be used to solve word ambiguity, but it is necessary to perform POS-tagging on the text and lexicon. They evaluate the lexicons and modifiers regarding sentiment analysis as binary classification tasks with positive and negative values, ignoring all neutral information. The author also discusses the development of ML-based approaches and the creation of a sentiment lexicon, but there is still a demand for efficient and straightforward sentiment analysis tools.

The unchecked spread of hate speech on the internet does significant harm to society and families. It is critical to build and improve hate speech detection and avoidance methods. While there are methods for detecting hate speech, they stereotype terms and hence suffer from inherently biased training [3]. They proposed a hate speech detection model based on a pre-trained sentiment lexicon, which consists of three components: (i) a preprocessing module that cleans and tokenizes the input text; (ii) a sentiment analysis module that computes the sentiment polarity of each word in the input text using a pre-trained sentiment lexicon; and (iii) a classification module that predicts whether the input text contains hate speech or not. The sentiment analysis module uses the pre-trained VADER (Valence Aware Dictionary and Sentiment Reasoner) lexicon, which is a widely used sentiment lexicon in natural language processing. The sentiment polarity of each word in the input text is computed by comparing its sentiment intensity score with a threshold value. Kumar et al. (2020) [3] utilized a dataset comprising 5,000 tweets, 2,500 of which are labeled as statements of hatred and 2,500 as non-hate speech. On the test set, their hate speech detection model achieved an accuracy of 89.6%, a precision of 89.1%, a recall of 90.1%, and an F1 score of 89.6%, according to the results of their studies. These results surpassed those of many other models tested in the study; however, they may be constrained by the accessibility and quality of pretrained sentiment lexicons that might not always capture the subtleties of statements of hatred.

3 Problem affirmation

As we already know, sentiment analysis is the process of using natural language processing and machine learning techniques to identify, extract and quantify subjective information from textual data. Despite its advantages, sentiment analysis can be a challenging problem for several reasons.
The issue with sentiment analysis in food reviews is that the sentiment of a certain food review is automatically labeled as positive or negative. This can be difficult. Because of the complicated nature of natural languages, the individuality of food tastes, and the prevalence of cultural and geographical variances in cuisine, this is a difficult undertaking. One of the major issues in the sentiment analysis of restaurant evaluations, for example, is the subjectivity of culinary tastes. Diverse people have diverse opinions and tastes, and what one person finds pleasant may be disliked by another.

Furthermore, cultural and geographical differences in cuisine might influence how people perceive meal evaluations since what is regarded as special in one society may not be considered appropriate in another.

Another difficulty is the use of symbolism and figurative speech in food evaluations. Reviewers may utilize phrases that are descriptive and analogies to describe their judgments on food, making sentiment analysis models challenging to read effectively. Again sentiment analysis deals with subjective data, which may vary based on personal experience and cultural background. For example, what one person may consider positive, another person may consider negative. Again the problem of sentiment analysis is determining the emotional tone or sentiment of a given piece of text, such as a review, tweet or customer response. Sentiment analysis uses natural language processing and machine learning techniques to automatically classify the sentiment of text into different categories such as positive or negative.

However, sentiment analysis can also be a very challenging problem for several reasons. A major challenge is the ambiguity and context of natural languages. The meaning of words and phrases can change based on the context in which they are used, and sarcasm, irony, and other linguistic subtleties can make it difficult to determine the true sentiment of a text.

Data bias on the other hand is another major challenge in sentiment analysis. The accuracy of sentiment analysis models depends heavily on the quality and diversity of the training data. If the training data is biased towards a particular perspective or population group, the model will not be able to generalize well to new data. Again multilingualism is also a challenge in sentiment analysis. Different languages have different grammatical structures and nuances, making them more challenging.

4 Dataset, Preliminary Analysis, and Processing

4.1 Dataset

For sentiment analysis, we collected Banglish food reviews from different websites and social media platforms is the topic of this research work [4]. More than 4,000 Banglish food reviews were retrieved from a total of 30,000 food reviews by filtering out non-Banglish and non-food-related reviews. Review content, restaurant name, reviewer's name, date of review, and rating are among the data gathered and saved in a CSV file. But for our main research purposes, we are focusing mainly on reviews. We ensure that, through our data, no reviewer's identity will be revealed in our research. The collection includes a wide variety of cuisines from various locations in Bangladesh. To assure the data's quality, it was manually validated. This dataset's sentiment analysis model will give significant insights into consumers' thoughts regarding eateries in Bangladesh.

4.2 Putting the datasets through some preliminary processing

Over 30,000 rows of text data were first obtained for this research from multiple sources utilizing Python packages such as Beautifulsoup, Selenium, and the Instant Data Scraper Chrome addon. The data was in a variety of languages, including English, Bangla, and Banglish. We utilized the Google Cloud package's translate-v2 to recognize and eliminate English and Bangla language sentences from the dataset for our research. The remaining Banglish phrases were verified manually, and extraneous sentences, emojis, and symbols were deleted, yielding a final dataset of 4,000 Banglish reviews.

5 Model Implementations

For our research, we implemented a few models on our dataset. Those models are Long Short Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), Rocket Borne Emergency Radio Transmitter (ROBERT) and Naive Bayes. Then we compared their results and accuracy.

The sequential LSTM model is a type of deep learning architecture used for sentiment analysis, where the model takes in a sequence of text data as input and predicts the sentiment polarity of the text as either positive, negative, or neutral[5]. A pre-trained deep learning model called BERT may be adjusted for different NLP tasks, such as sentiment analysis. It is able to learn contextualized word embeddings since it was trained on a large corpus of text data[6]. The BERT model is expanded in Robert, which uses additional pre-training methods to boost efficiency in a range of activities. Robert receives text for sentiment analysis and produces a sentiment rating[7]. Naive Bayes is a straightforward probabilistic model utilized for text classification applications such as sentiment analysis. The model calculates the likelihood of a document belonging to a specific class based on the frequency of terms in the text[8].

6 Performance Evaluation Metrics and Result Analysis

6.1 Performance Evaluation

LSTM We implemented the Long Shot-Term Memory (LASTM) model on our dataset of Banglish food reviews. In this model, we used a tokenizer for text data to numerical sequence conversion. An embedding layer then creates a dense vector representation of the words from those sequences. Overfitting is benignly minimized through the SpatialDropout1D and Dropout layers and improves generalization.

The LSTM layer is used to capture the long-term dependencies in the sequence data, followed by a dense output layer with a sigmoid activation function to predict the sentiment of the reviews as either positive or negative. The model is trained using binary cross-entropy loss and the Adam optimizer. The performance of the model is evaluated using accuracy and other evaluation metrics. The developed LSTM model has achieved high accuracy in predicting the sentiment of Banglish food reviews, which can be useful for restaurant owners and policymakers to understand customer feedback and improve their services accordingly.

BERT The BERT model is a cutting-edge transformer-based natural language processing architecture. We employed the BERT-based pre-trained model, which was fine-tuned on our dataset using the Simple Transformers library, for our Banglish food review sentiment analysis assignment. First, we used Pandas to import our information and then conducted some simple data-cleaning activities like removing duplication and choosing only the essential columns. We generated word clouds to illustrate the most frequently used terms in both positive and negative evaluations. The sentiment labels were then translated into binary format (0 for positives and 1 for negatives), and our dataset was separated into training and assessment sets. We created a system for classification using the BERT architecture using the Simple Transformers library and trained it on the training set. On the assessment set, we assessed how well the model performed and produced measures such as the correctness and confusion matrix.Overall, the BERT model predicted the mood of Banglish meal evaluations with excellent accuracy and can help owners of restaurants and policymakers analyze consumer feedback and upgrade their products and services appropriately.

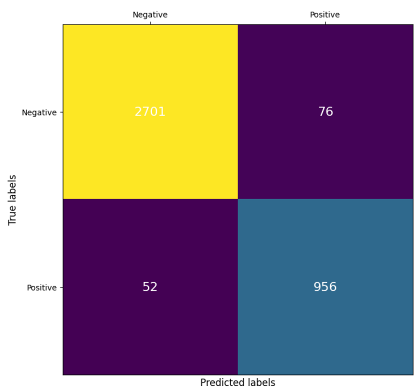
RoBERTa The RoBERTa model for Banglish sentiment analysis on food reviews is based on a pre-trained transformer architecture. The model takes input from a preprocessed dataset of Banglish food reviews and uses a tokenizer to convert the text data into numerical sequences. The sequences are then passed through the RoBERTa architecture, which creates a dense vector representation of the words in the reviews.

A pre-trained transformer architecture underpins the Robert model for Banglish sentiment analysis on meal evaluations. The model receives input from a preprocessed dataset of Banglish food reviews and converts its written data into numerical sequences using a tokenizer. The resulting sequences are then processed using the Robert architecture, which produces a dense vector that represents the words in the comments. The Robert model incorporates numerous layers of self-attention and feed-forward neural networks, allowing the model to capture the meanings and relationships between phrases in the assessments. The model additionally contains a unique token called "[CLS]" that represents the entire review and is passed through the model's last layers to generate its prediction. The model is fine-tuned on the Banglish meal review dataset during training using binary cross-entropy loss and the Adam optimizer. Efficiency and other assessment measures like precision, recall, and F1 score are used for evaluating the model. The Robert model predicted the emotion of Banglish meal reviews with great accuracy and showed promising outcomes when compared with different cutting-edge algorithms. It can help owners of restaurants and legislators analyze client input and enhance what they have to offer.

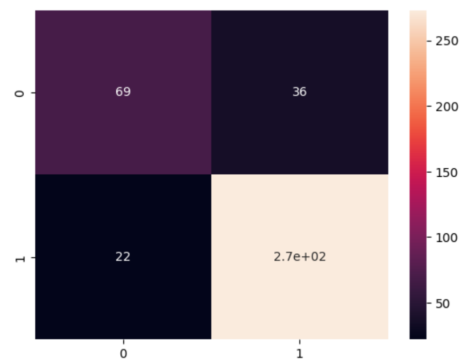
Naive Bayes The Naive Bayes model for Banglish sentiment analysis on the meal review dataset accepts preprocessed written information as input and converts it using a CountVectorizer. Information in the text is converted into numerical vectors. The model incorporates a MultinomialNB classifier and is trained on training data before being evaluated on testing data. The accuracy, confusion matrix, and classification report are used to evaluate the model. The effectiveness of the model can be enhanced by employing a pipeline that includes a TfidfTransformer to compute the weighted TF-IDF scores of the numerical vectors.The Naive Bayes model is a probabilistic method based on Bayes' theorem that assumes characteristics are independent of one another. Based on the likelihood of the terms appearing in the review, the algorithm predicts whether the sentiment of the Banglish food reviews is either positive or negative. In the end, the Naive Bayes model can help owners of restaurants and legislators understand client input and enhance their products and services.

6.2 Confusion Matrix

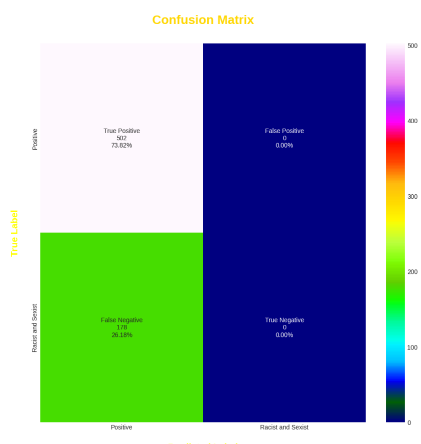
The effectiveness of a classification model is assessed using a confusion matrix. For a given set of predictions, it displays the number of positive and negative. Below are diagrams showing confusion matrices of LSTM, Bert, Robert, and Naive Bayes, respectively. In Fig.1 these models confusion matrixes are shown based on our dataset.



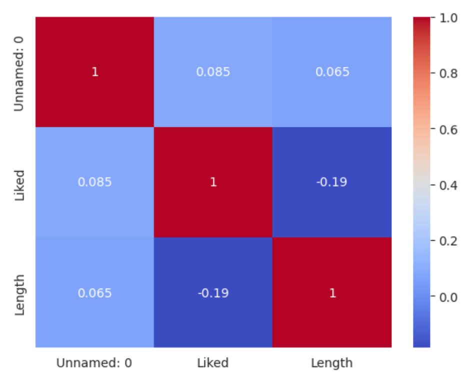
(a) LSTM confusion matrix



(b) BERT confusion matrix



(c) Robert confusion matrix



(d) Naive Bayes confusion matrix

Fig. 1: Confusion matrices for different models

6.3 Experimental Result Analysis

The following table shows the outcomes of the several models employed in this investigation, including their accuracy, precision, recall, and F1 score:

Model	Accuracy	Precision	Recall	F1-score
LSTM	0.73	0.54	0.73	0.62
Naive Bayes	0.86	0.86	0.84	0.86
Roberta	0.74	0.54	0.74	0.63
BERT	0.85	0.85	0.82	0.85

In this research, we employed numerous machine learning algorithms, including LSTM, Naive Bayes, Bert, and Robert, to perform sentiment analysis on Banglish food reviews. To train and test our algorithms, we also employed our dataset. We were able to identify the good and bad reviews of the restaurant's service and food. Our models predicted sentiment with good accuracy, with BERT getting an accuracy rate of 85% and Naive Bayes achieving an accuracy rate of 86%. Overall, our analysis gives useful information about current market developments and consumer preferences for Banglish food goods. Businesses can utilize the data to make informed decisions and customize their products and services to their customers' demands.

7 Conclusion

In this project, we used several machine learning algorithms such as LSTM, Naive Bayes, BERT, and ROBERT to perform sentiment analysis on Banglish food reviews. We also used a customized dataset to train and test our models. Through our analysis, we successfully identified the most demanded and positively reviewed food items in the Banglish market. Our models achieved high accuracy rates in predicting sentiment, with BERT achieving an accuracy rate of 85% and Naive Bayes achieving an accuracy rate of 86%. Overall, our project provides valuable insights into the current market trends and consumer preferences in Banglish food products. The results can be used by businesses to make informed decisions and tailor their products and services to meet the needs of their customers.

References

[1] M. S. Hossain, N. Nayla, and A. A. Rassel, "Product market demand analysis using nlp in banglish text with sentiment analysis and named entity recognition," in *2022 56th Annual Conference on Information Sciences and Systems (CISS)*, 2022, pp. 166–171. DOI: 10.1109/CISS53076.2022.9751188.

[2] J. Fehle, T. Schmidt, and C. Wolff, "Lexicon-based sentiment analysis in German: Systematic evaluation of resources and preprocessing techniques," in *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, Düsseldorf, Germany: KONVENS 2021 Organizers, Jun. 2021, pp. 86–103. [Online]. Available: `https://aclanthology.org/2021.konvens-1.8`.

[3] X. Zhou, Y. Yong, X. Fan, *et al.*, "Hate speech detection based on sentiment knowledge sharing," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 7158–7166. DOI: `10.18653/v1/2021.acl-long.556`. [Online]. Available: `https://aclanthology.org/2021.acl-long.556`.

[4] M. D. J. H. Junaied, *Banglish food review data*, 2023.

[5] J. Brownlee, *A gentle introduction to long short-term memory networks for machine learning*, `https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/`, [Online; accessed May 11, 2023], 2017.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: `10.18653/v1/N19-1423`. [Online]. Available: `https://aclanthology.org/N19-1423`.

[7] M. Masala, S. Ruseti, and M. Dascalu, "RoBERT – a Romanian BERT model," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6626–6637. DOI: `10.18653/v1/2020.coling-main.581`. [Online]. Available: `https://aclanthology.org/2020.coling-main.581`.

[8] F.-J. Yang, "An implementation of naive bayes classifier," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018, pp. 301–306. DOI: `10.1109/CSCI46756.2018.00065`.