TASK-6

Banglish food review sentiment analysis.

by

MD Junaied Hossain
20101204
Mohammad Tanjil Islam
15301110
MD. Fardin Ahsan
20101208

Department of Computer Science and Engineering
Brac University
September 2022

Table of Contents

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, 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

Chapter 1

Introduction

1.1 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 sentiment. 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.

1.2 Research Objectives

The objective of our research is to evaluate and compare the performance of various NLP and machine learning models on our dataset for sentiment analysis, to identify the most accurate and efficient model that can predict the sentiment of Banglish sentences, whether it is positive or negative. For our model, we will train it on our own collected dataset. Before that, we will process our data to remove unwanted elements from it. Once we have our dataset, we will train and test various models on it. Through an iterative process of training and testing, we would like to fix all the bugs and errors and compare which models are performing well on our dataset.

1.3 Problem Statement

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 problem with sentiment analysis in food reviews is that the sentiment of a given food review is automatically classified into different categories as positive, negative or neutral. This can be a challenging task due to the complexity of natural languages, the subjectivity of food preferences, and the presence of cultural and regional variations in food. For example:
One of the main challenges in sentiment analysis of food reviews is the subjectivity of food preferences. Different people have different tastes and preferences, and what one person may find tasty, another person may dislike. Additionally, cultural and regional differences in food can also affect the perception of food reviews, as what is considered a delicacy in one culture may not be as accepted in another culture. Another challenge is the use of figurative language and metaphors in food reviews. Reviewers may use descriptive language and comparisons to express their opinions about food, making it difficult for sentiment analysis models to accurately interpret sentiment. 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, negative or neutral. 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.

Chapter 2

Literature Review

Md Sabbir Hossain[3], discusses the importance of product market demand analysis for business strategies, particularly 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. The authors 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 percent in Spacy Custom Named Entity Recognition, 95.51 percent in Amazon Comprehend Custom NER, and 87.02 percent 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. The paper discusses the challenges of collecting and labelling the Banglish text data set and the importance of sentiment analysis and natural language processing for obtaining, quantifying, and analyzing consumer preferences. Sentiment analysis is also used to observe consumers' preferences, desired models, and brands, and is a useful tool for entrepreneurs to originate efficient business strategies

Jakob Fehle[1], represents lexicon-based sentiment analysis in German using systematic evaluation of resources and preprocessing techniques. They analyzed 20 sentiment-annotated corpora and 19 sentiment lexicon resources for German texts from various topics. The paper talks about a subfield of affective computing called sentiment analysis" that focuses on identifying and analyzing human sentiment and emotions across a range of application domains. 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. The two main branches of sentiment analysis techniques are lexicon-based (also known as rule-based or dictionary-based methods) and machine learning (ML)-based methods. They perform two steps to clean the texts of all corpora before evaluation, and it is observed by the author that the best combination of modifiers, and different methods are cross-evaluated and compared based on classification metrics. 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 used a variety of digital libraries and search engines to look for appropriate corpora and lexicons for sentiment analysis in German, such as the ACM Digital Library, the ACL Anthology, the IEEE, Springer Verlag, and KONVENS. They evaluate the lexicons and modifiers regarding sentiment analysis as binary classification tasks with positive and negative values, ignoring all neutral information. Before delving deeper into the most effective lexicon-modifier combinations, they show lexicon performance without the use of modifiers as well as outcomes based on modifier usage. 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.

Kumar et al. (2020)[2], proposed a hate speech detection model based on a pre-trained sentiment lexicon. The model consists of three components: (1) a preprocessing module that cleans and tokenizes the input text, (2) a sentiment analysis module that computes the sentiment polarity of each word in the input text using a pre-trained sentiment lexicon, and (3) 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 VADER lexicon contains words and phrases that are associated with positive or negative sentiment, as well as a sentiment intensity score for each word or phrase. The sentiment polarity of each word in the input text is computed by comparing its sentiment intensity score with a threshold value. To train and evaluate their hate speech detection model, Kumar et al. (2020) used a dataset of tweets that were manually annotated for hate speech. The dataset contains 5,000 tweets, of which 2,500 are labelled as hate speech and 2,500 are labelled as non-hate speech. The results of their experiments showed that their hate speech detection model achieved an accuracy of 89.6 percent, a precision of 89.1 percent, a recall of 90.1 percent, and an F1 score of 89.6 percent on the test set. These results outperformed those of several other models that were compared in the study, including a baseline model that used bag-of-words features and a logistic regression classifier, and a model that used pre-trained word embedding. However, the approach may be limited by the availability and quality of pre-trained sentiment lexicons, which may not always capture the nuances of hate speech.

Chapter 3

Data Collection

3.1 Data Collection

This research paper focuses on sentiment analysis of Banglish food reviews collected from various websites and social media platforms. Out of a total of 30,000 food reviews, over 5,000 Banglish food reviews were extracted by filtering out non-Banglish and non-food related reviews. The collected data, stored in a CSV file, includes review text, restaurant name, reviewer's name, date of review, and rating. The dataset covers a diverse range of cuisine types from different regions of Bangladesh. The data were manually checked to ensure their quality. The sentiment analysis model developed using this dataset will provide valuable insights into customers' opinions about restaurants in Bangladesh.

3.2 Data Pre-processing

For this research, over 30,000 rows of text data were initially collected from various sources using Python libraries such as Beautifulsoup, Selenium, and Instant Data Scraper Chrome extension. The data contained mixed languages, including English, Bangla, and Banglish. To filter the dataset for our research, we used the Google Cloud package's translate-v2 to detect and remove English and Bangla language sentences. The remaining Banglish sentences were manually checked and irrelevant sentences, emojis, and symbols were removed, resulting in a final dataset of 5,000 Banglish reviews. The dataset was manually labelled, and 2,777 reviews were identified as negative while the remaining reviews were considered positive. This processed dataset will be used to train and evaluate sentiment analysis models for Banglish food reviews.

Chapter 4

Research Methodology

4.1 Workflow of proposed model

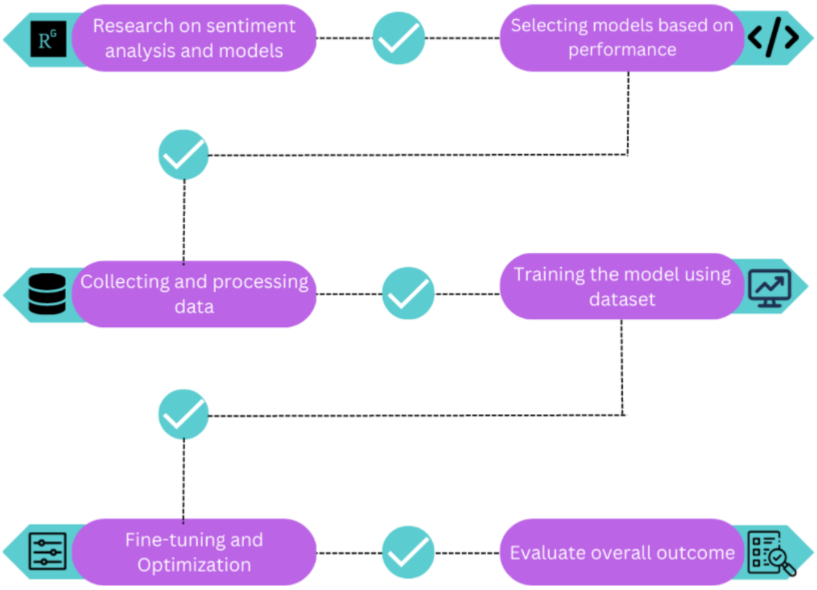
Research on sentiment analysis and models
Selecting models based on performance
Collecting and processing data
Training the model using dataset
Fine-tuning and Optimization
Evaluate overall outcome

Figure 4.1: Flowchart for the proposed work plan

Our research plan Figure: 3.1, involves studying sentiment analysis models and selecting the best-performing one. We will collect and process relevant data and use it to train the selected model. We will then fine-tune and optimize the model for better accuracy and efficiency. Finally, we will evaluate the overall outcome of our study to determine the effectiveness of the model in analyzing sentiment.

4.2 Approaches

4.2.1 Model Description

LSTM

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. The LSTM units in the model can selectively remember or forget information based on the input sequence, and the output of the final layer is used to make the prediction. The model is trained using backpropagation through time, which updates the weights and biases of the model to minimize the loss between the predicted sentiment and the actual sentiment label. Overall, the sequential LSTM model is a powerful tool for sentiment analysis due to its ability to capture the complex relationships between words in a text sequence.

BERT

BERT is a pre-trained deep learning model that can be fine-tuned for various NLP tasks, including sentiment analysis. It is trained on a large corpus of text data, allowing it to learn contextualized word embeddings. For sentiment analysis, BERT takes in a piece of text and outputs a sentiment score. This makes BERT a powerful tool for sentiment analysis, as it can capture the complex meaning and context of language to provide more accurate results.

RoBERTa

RoBERTa is a pre-trained transformer-based deep learning model that can be fine-tuned for various NLP tasks, including sentiment analysis. RoBERTa is an extension of the BERT model that incorporates additional pre-training techniques to improve its performance on a variety of tasks. For sentiment analysis, RoBERTa takes in a piece of text and outputs a sentiment score. RoBERTa has shown strong performance on many NLP tasks, including sentiment analysis, due to its ability to capture the contextual meaning of words and phrases in natural language text.

Naive Bayes

Naive Bayes is a simple probabilistic model used for text classification tasks, including sentiment analysis. The model works by calculating the probability of a document belonging to a particular class based on the frequency of words in the document. The Naive Bayes algorithm makes the assumption that the presence of each word in a document is independent of the presence of other words, hence the term ”naive”. Despite this simplifying assumption, Naive Bayes can be very effective for sentiment analysis tasks due to its ability to handle large datasets and its fast training and prediction times.

4.2.2 Model Analysis

LSTM

The model for Banglish sentiment analysis is based on a Long Short-Term Memory (LSTM) 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 an embedding layer, which creates a dense vector representation of the words in the reviews. The model includes SpatialDropout1D and Dropout layers to prevent overfitting and improve 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.

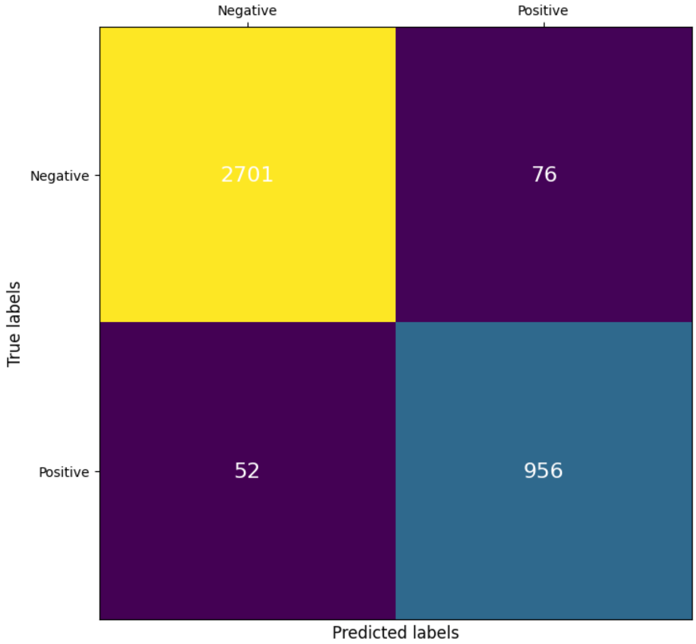

Figure 4.2: LSTM confusion matrix

BERT

The BERT model is a state-of-the-art transformer-based architecture for natural language processing. For our Banglish food review sentiment analysis task, we used the BERT-base-cased pre-trained model which was fine-tuned on our dataset using the SimpleTransformers library. First, we loaded our dataset using pandas and performed some basic data cleaning tasks such as dropping duplicates and selecting only the relevant columns. We also created word clouds to visualize the most common words in the positive and negative reviews. We then split our dataset into training and evaluation sets and converted the sentiment labels into binary format (0 for positive and 1 for negative). We used the SimpleTransformers library to create a classification model with the BERT architecture and trained it on the training set. We evaluated the model's performance on the evaluation set and computed various metrics such as accuracy and confusion matrix. Overall, the BERT model achieved high accuracy in predicting the sentiment of Banglish food reviews and can be useful for restaurant owners and policymakers to understand customer feedback and improve their services accordingly.

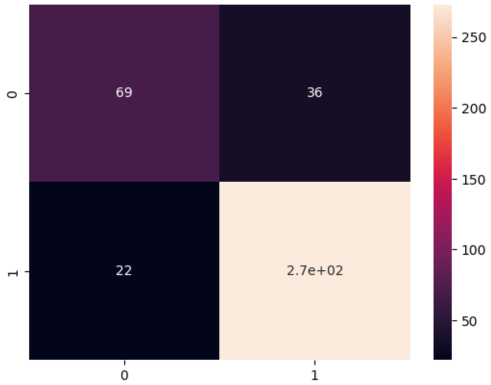

Figure 4.3: BERT confusion matrix

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. The RoBERTa model includes multiple layers of self-attention and feed-forward neural networks that allow the model to capture the context and dependencies between the words in the reviews. The model also includes a special token called the "[CLS]" token, which is used to represent the entire review and is passed through the model's final layers to make the prediction. During training, the model is fine-tuned on the Banglish food review dataset using binary cross-entropy loss and the Adam optimizer. The model is evaluated using accuracy and other evaluation metrics such as precision, recall, and F1 score.
The RoBERTa model has achieved high accuracy in predicting the sentiment of Banglish food reviews and has shown promising results in comparison to other state-of-the-art models. It can be useful for restaurant owners and policymakers to understand customer feedback and improve their services accordingly.

Naive Bayes

The Naive Bayes model for Banglish sentiment analysis on the food review dataset takes input from preprocessed text data and uses a CountVectorizer to convert the text data into numerical vectors. The model includes a MultinomialNB classifier and is trained using the training data and tested using the testing data. The model is evaluated using accuracy, confusion matrix, and classification report. The

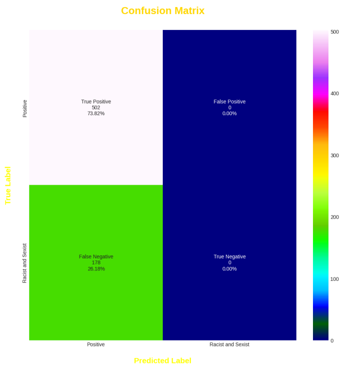

Figure 4.4: Robert confusion matrix

performance of the model can be improved by using a pipeline that includes a Tfidf-Transformer to calculate the weighted TF-IDF scores of the numerical vectors. The Naive Bayes model is a probabilistic algorithm based on Bayes' theorem, which assumes that the features are independent of each other. The model is used to predict the sentiment of the Banglish food reviews as either positive or negative based on the probability of the occurrence of the words in the review. Overall, the Naive Bayes model can be a useful tool for restaurant owners and policymakers to analyze customer feedback and improve their services.

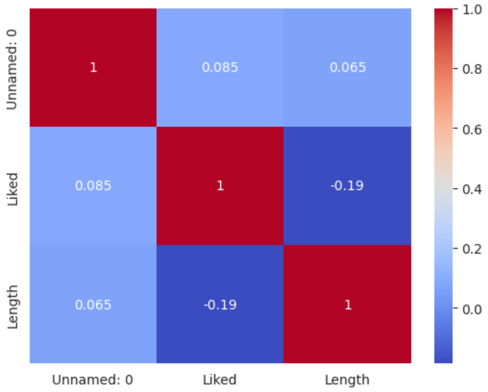

Figure 4.5: Naive Bayest confusion matrix

Chapter 5

RESULT and Accuracy

5.1 RESULT and Accuracy

The table below summarizes the results of the different models used in this study, 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

Based on the results and accuracy, it appears that the Naive Bayes model performed the best, achieving an accuracy of 86% and relatively high precision, recall, and F1-score values. The BERT model also had relatively high accuracy and precision but lower recall and F1-score values. The LSTM and Roberta models had lower accuracy and precision values, with Roberta, in particular, performing poorly in terms of precision and recall. However, it is important to note that the metrics are calculated for each model's performance on a different set of labels or classes, so further analysis and evaluation would be necessary to determine the best model for a specific task.

Chapter 6

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.

Bibliography

[1] 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.

[2] 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.

[3] 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.