REPORT

BENGALI NEWS ARTICLES SENTIMENT ANALYSIS

ANTARA DAS [IIITD]

INTRODUCTION

Sentiment analysis, a vital domain within natural language processing (NLP), categorises the sentiments conveyed in a text by evaluating its polarity, which can be positive, negative, or neutral. This work is designed to predict the sentiment of a Bengali news article and analyse a collection of such predicted data. The system is expected to be efficiently implemented for real-time inference.

PROBLEM STATEMENT

The system will periodically query a designated directory containing Bengali (or any other low-resource regional Indian language) news articles at regular intervals to predict the sentiment of each new article added to the directory and store those predicted results in a designated output directory. The program should be capable of concurrently processing multiple articles in a multi-threaded manner. Analysing the predicted results through different data analysis and visualisation techniques is also required. The use of publicly available pre-trained models and libraries should be preferred to perform this task.

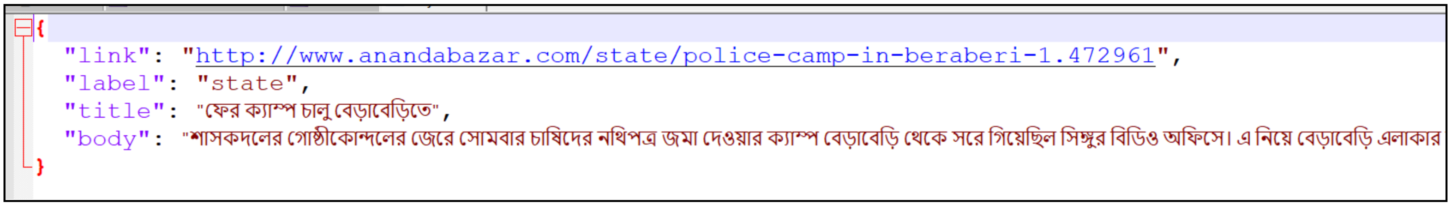
TOOLS USED

- Deep Learning Library (PyTorch, transformers)
- ML Libraries (NumPy, Pandas, scikit-learn)
- NLP Tools (BNLTK, BNLP, WordCloud)
- Data Visualization Libraries (matplotlib, seaborn, plotly)

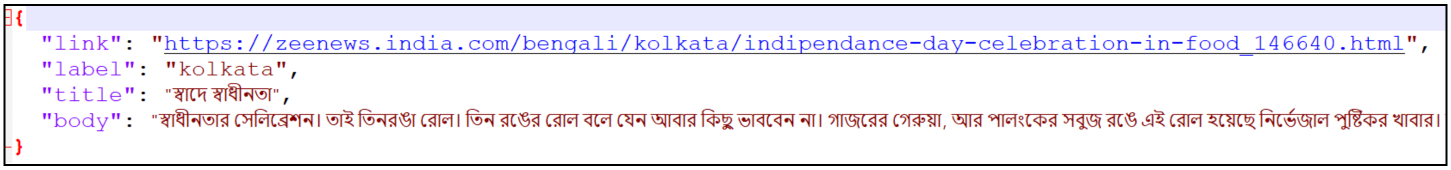
DATASET

A dataset of Bengali news articles collected from the websites of two newspapers (Zeenews and Anandabazar) is publicly available here. This data contains 7411 news articles and the news title and category (e.g., sports, entertainment, national, etc.). In this work, this dataset is used for user inference or checking the desired system functionality and the analysis of the predictions.

A sample data file is given as follows:



```
{
   "link": "http://www.anandabazar.com/state/police-camp-in-beraberi-1.472961",
   "label": "state",
   "title": "ফের ক্যাম্প চালু বেড়াবেড়িতে",
   "body": "শাসকদলের গোষ্ঠীকোন্দলের জেরে সোমবার চাষিদের নথিপত্র জমা দেওয়ার ক্যাম্প বেড়াবেড়ি থেকে সরে গিয়েছিল সিঙ্গুর বিডিও অফিসে। এ নিয়ে বেড়াবেড়ি এলাকার
}
```



```
{
  "link": "https://zeenews.india.com/bengali/kolkata/indipendance-day-celebration-in-food_146640.html",
  "label": "kolkata",
  "title": "স্বাদে স্বাধীনতা",
  "body": "স্বাধীনতার সেলিব্রেশন। তাই তিনরঙা রোল। তিন রঙের রোল বলে যেন আবার কিছু ভাববেন না। গাজরের গেরুয়া, আর পালংকের সবুজ রঙে এই রোল হয়েছে নির্ভেজাল পুষ্টিকর খাবার।
}
```

Due to the unavailability of any pre-trained model in the Bengali language, particularly for sentiment analysis, an existing BERT model, which was trained on a Bengali corpus for masked language modelling, was required to be finetuned. To perform this, a Bengali dataset with sentiment labels was required. However, the news articles dataset was not labelled on sentiments, and hence, there was a requirement to collect another dataset for this finetuning task. This sentiment classification dataset is collected from the data used in different related works[1, 2]. All these data contain the user comments on popular news articles in an online Bengali news portal, and each of them is labelled with its corresponding sentiments (Negative denotes 0, Positive denotes 1, and Neutral denotes 2).

METHODOLOGY

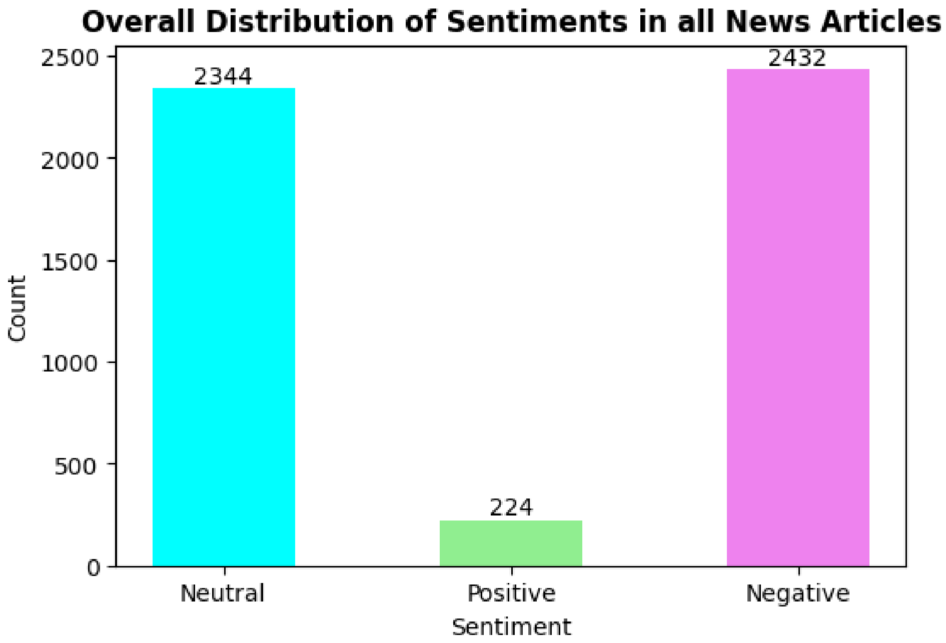
The first step involved here is data preparation. The sentiment-labelled datasets are combined together in a CSV file, and then they are preprocessed to remove stopwords, punctuations, URLs, emails, emojis, and digits, as these are irrelevant to the downstream task. Then, this dataset is divided into train and test splits containing 25000+ and 5000+ entries, respectively. The next task was to create individual JSON files for each news article from the collection of all articles.

A classical ML-based technique is tried out first to perform this sentiment classification task by training a Support Vector Machine. To obtain the word embedding of the input Bengali text, a pre-trained Fasttext library is used from the BNLP toolkit. A 100-dimensional vector is obtained using this non-contextual embedding technique, and this feature vector is used as training data in the Support Vector Classifier to predict the sentiment. Due to this method's low accuracy and F1 score, a DL-based technique is used for improvement.

Two popular LLMs are found in the Hugging Face model collection, pre-trained on a huge Bengali corpus from Wikitext on masked language modelling. These two models are finetuned for this downstream task of sentiment classification by adding a linear layer on top of the last hidden layer of the model. Training is performed on these last two layers, freezing all the lower layers.

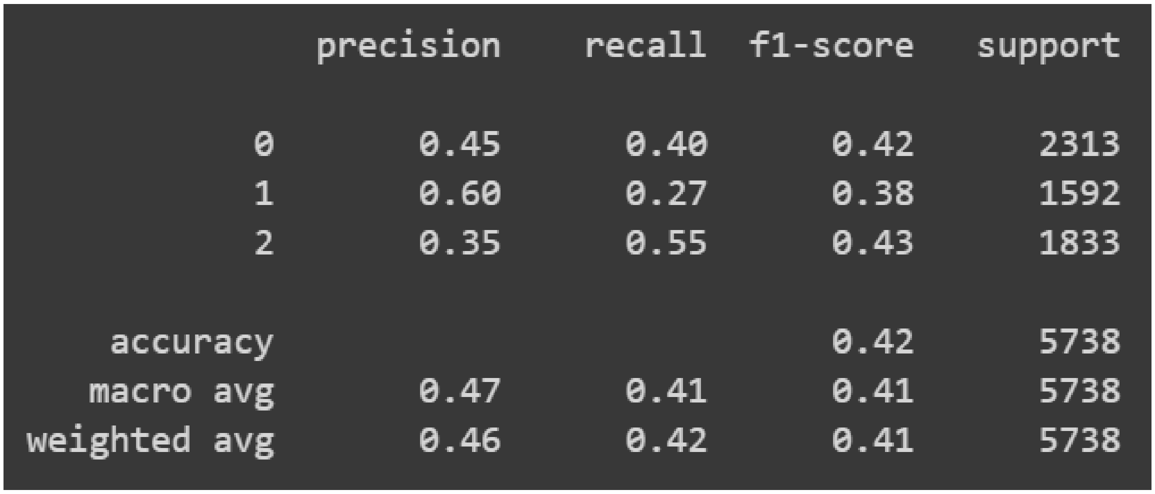
The last step was to design the code pipeline for user inference. This piece of code assumes the input file to be in JSON format, and the predicted output is stored as a JSON file with the same filename as the input file. It is also assumed that the system can support a maximum of five threads simultaneously to execute this task. A scheduler periodically checks if a new file is added to the directory containing all news articles every X minutes (user-defined parameter) and processes them in the multi-threaded code pipeline to generate the prediction files. The texts are preprocessed to remove unnecessary characters or stopwords before passing them to the model for inference. Also, an entire news article is divided into multiple sentences using a sentence tokenizer to ensure that those lengthy news articles do not surpass the maximum token length allowed for the model. The final prediction scores are an aggregation of scores of all sentences, and the class with the maximum score is determined as the predicted class. The output file contains the predicted sentiment, the predicted values of all possible sentiments (classes), the date and timestamp of the file generation, and the news category.

RESULT AND ANALYSIS

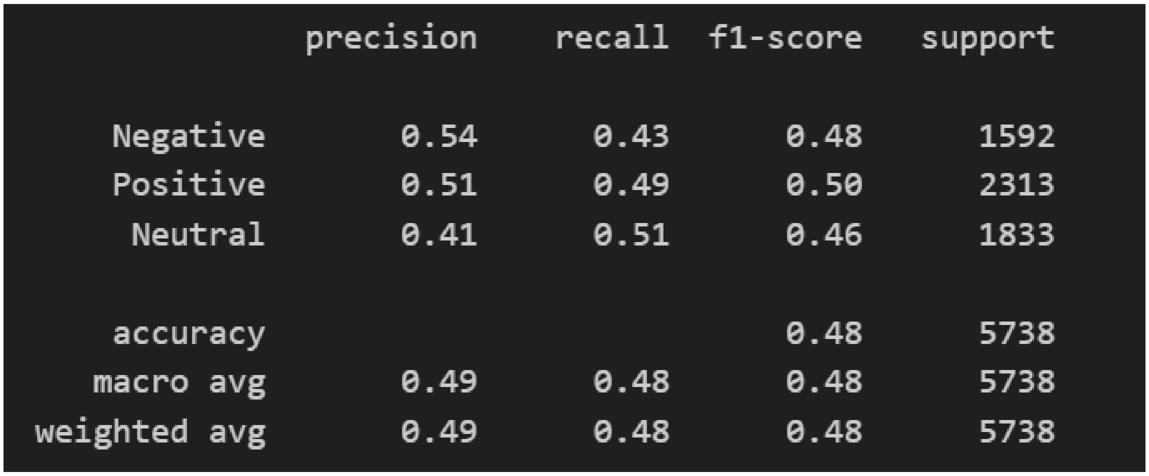


A sample of 5000 predicted files is used for the entire analysis task. From the overall distribution given above, it’s clear that positive sentiments are rarely present in the news. This could also indicate that the model is biased towards the neutral and negative classes. However, this is highly unlikely because the train and test dataset was well balanced over all classes.

The label-wise test performance metrics did not vary significantly, as shown below in the two performance reports: The first one is generated by the ML approach, and the BERT-based approach gives the second one:



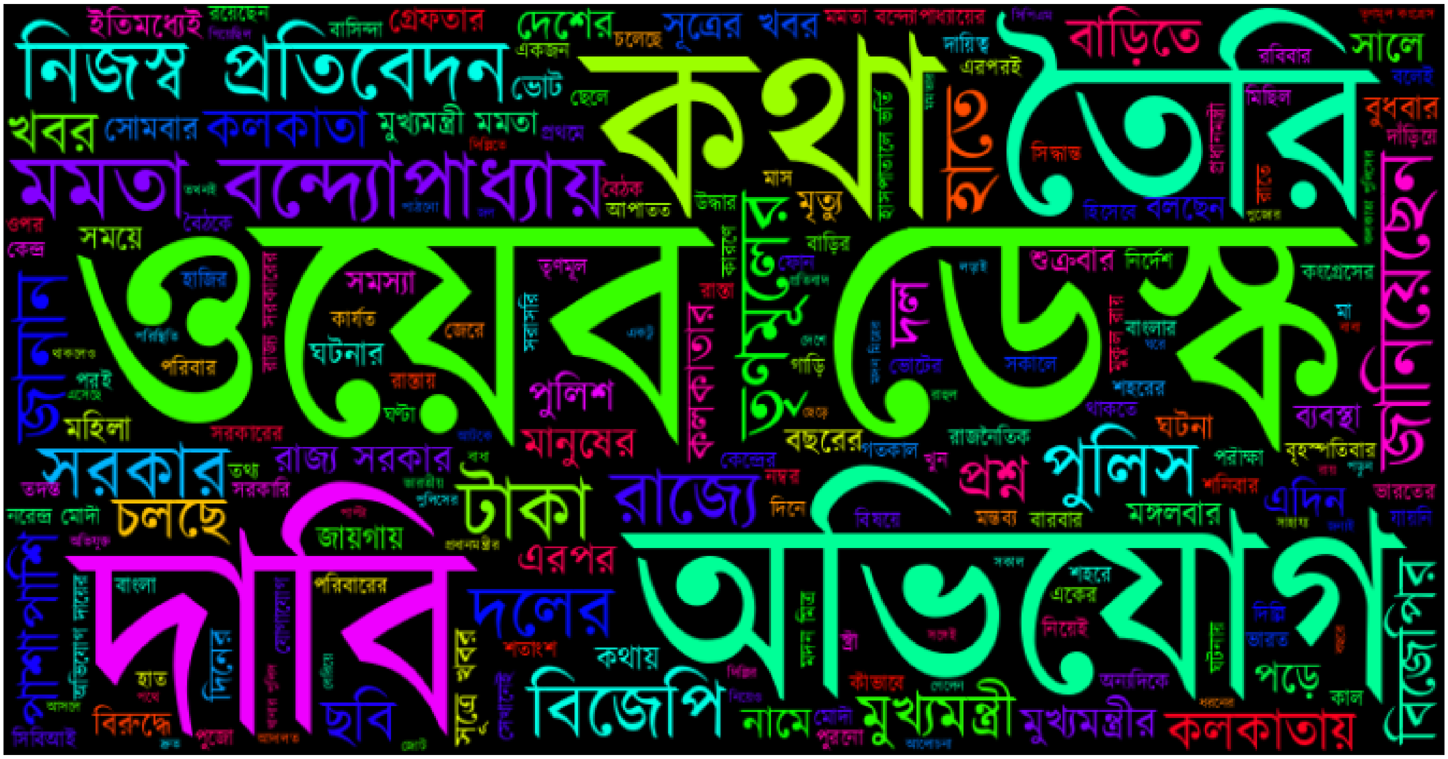
```
              precision    recall  f1-score   support

           0       0.45      0.40      0.42      2313
           1       0.60      0.27      0.38      1592
           2       0.35      0.55      0.43      1833

    accuracy                           0.42      5738
   macro avg       0.47      0.41      0.41      5738
weighted avg       0.46      0.42      0.41      5738
```



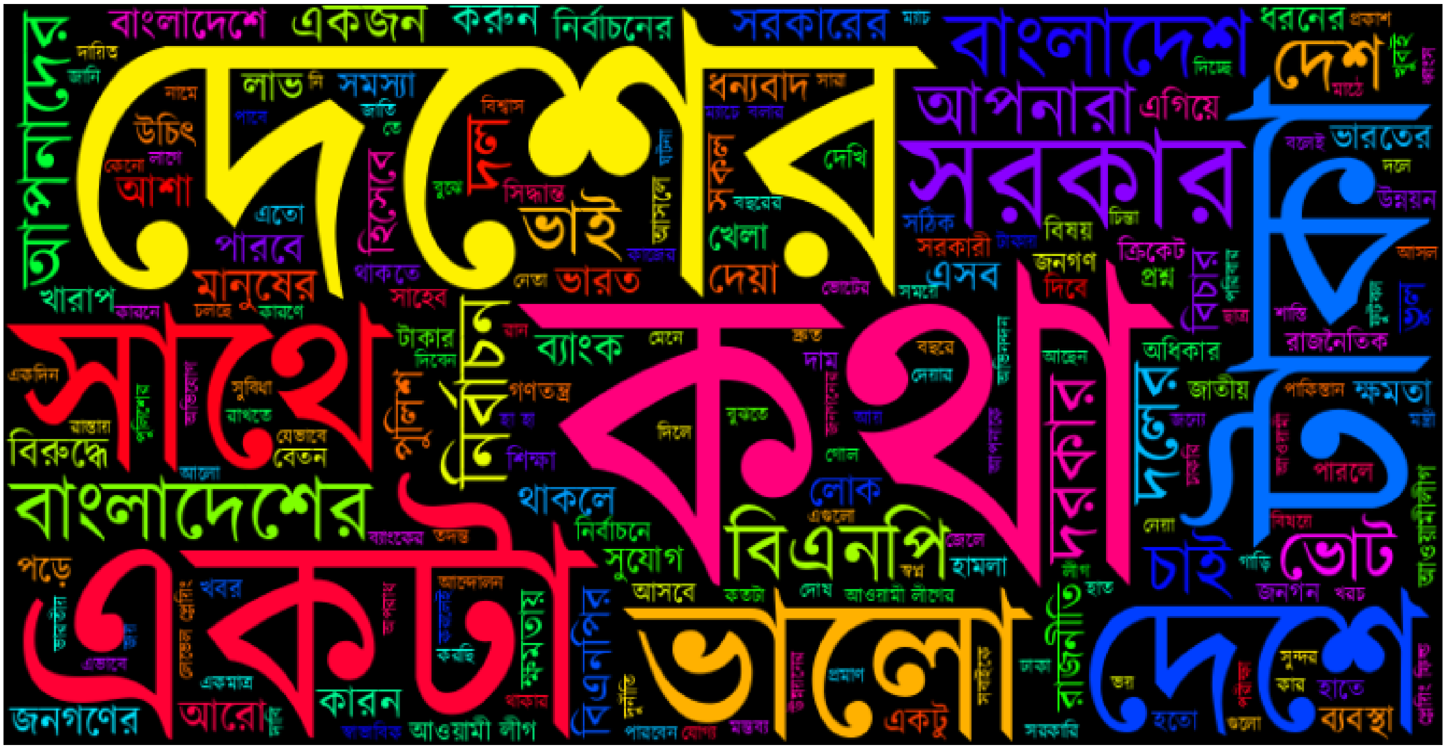
```
              precision    recall  f1-score   support

    Negative       0.54      0.43      0.48      1592
    Positive       0.51      0.49      0.50      2313
     Neutral       0.41      0.51      0.46      1833

    accuracy                           0.48      5738
   macro avg       0.49      0.48      0.48      5738
weighted avg       0.49      0.48      0.48      5738
```

This observation about the two datasets is also confirmed by the visualisation of the two datasets using word clouds, in which the most frequently occurring words are highlighted in an image.

In the word cloud of the news articles, positive words are not observed, whereas negative words are present (e.g., complain, murder, demand, against, etc.) in large fonts.

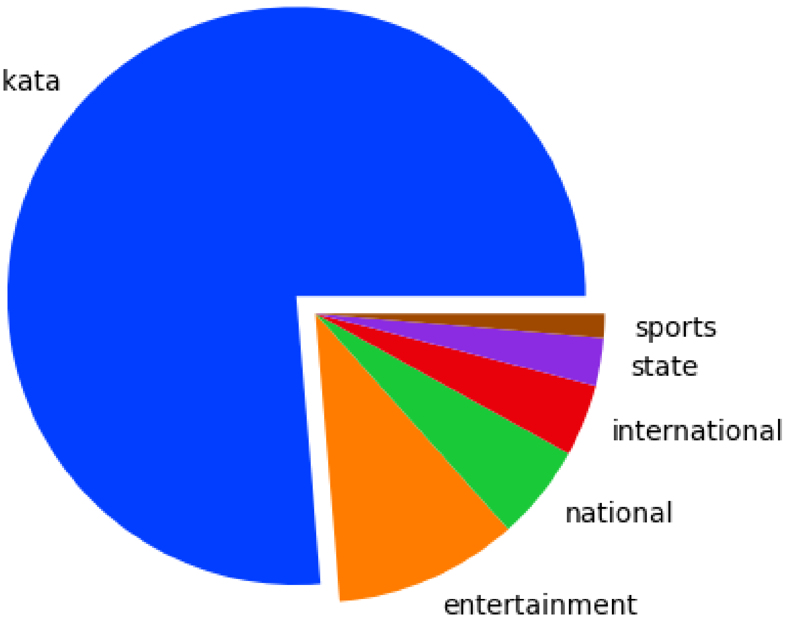


However, in the word cloud of the training dataset, positive (e.g., good, beautiful, opportunity, hope, thanks, peace, etc.) and a few negative words (e.g., bad, problem, etc.) are visible.

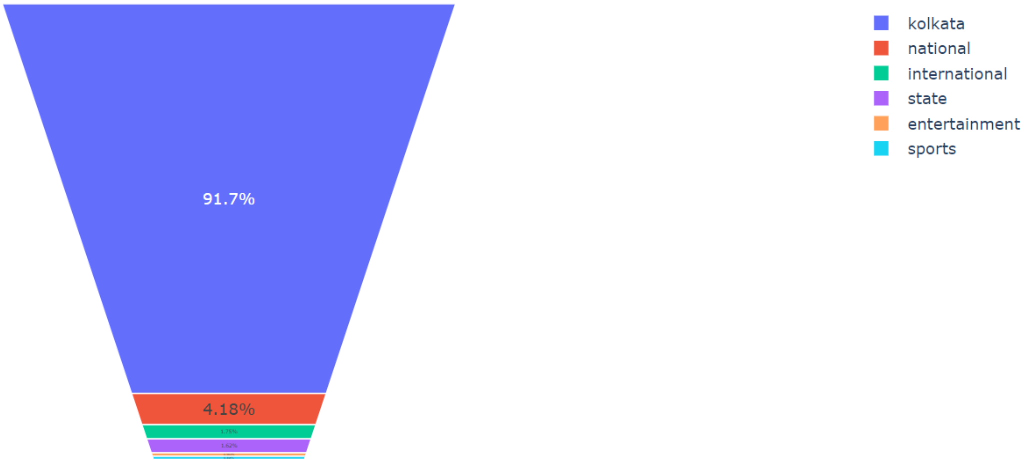


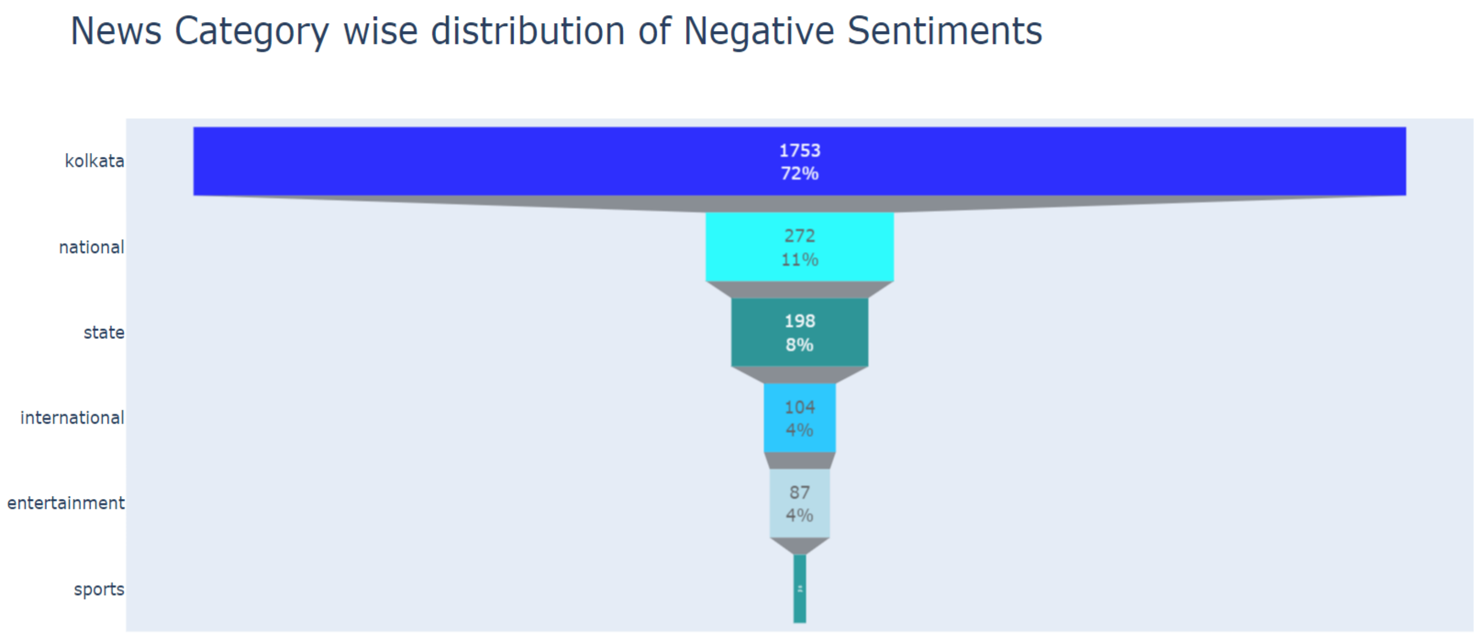
In the original dataset of news articles, those are tagged with different categories like state, sports, national, entertainment, etc. So, the category-wise analysis of the sentiments of news articles is given in the following pie and funnel charts:

News Category wise distribution of Positive Sentiments



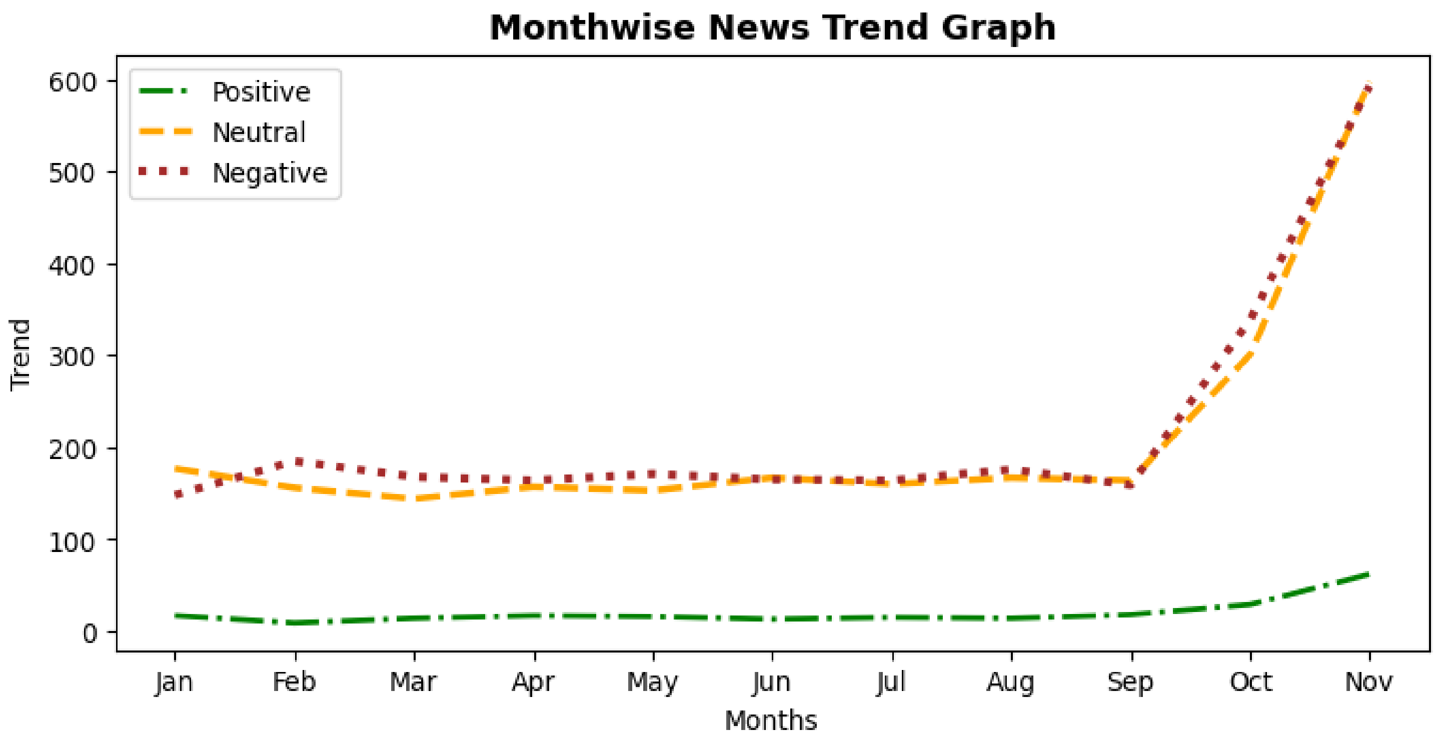
News Category wise distribution of Neutral Sentiments





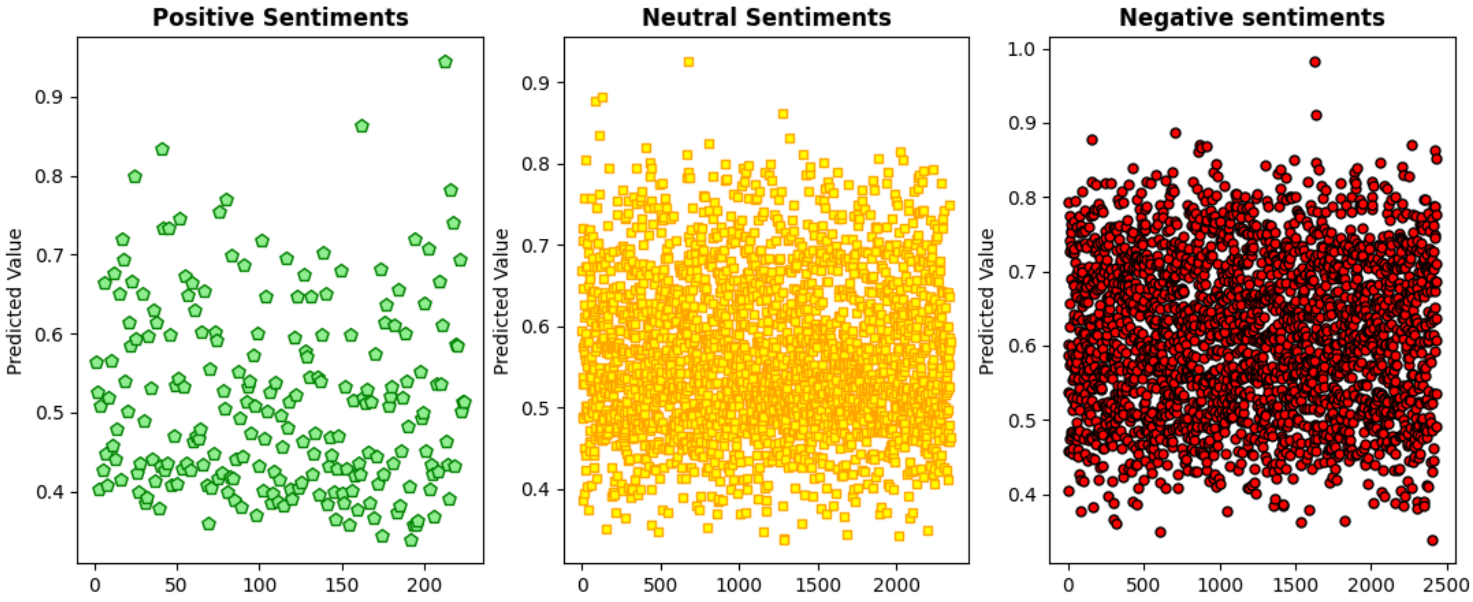
News Category wise distribution of Negative Sentiments

A month-wise trend graph of the sentiments is plotted as follows:

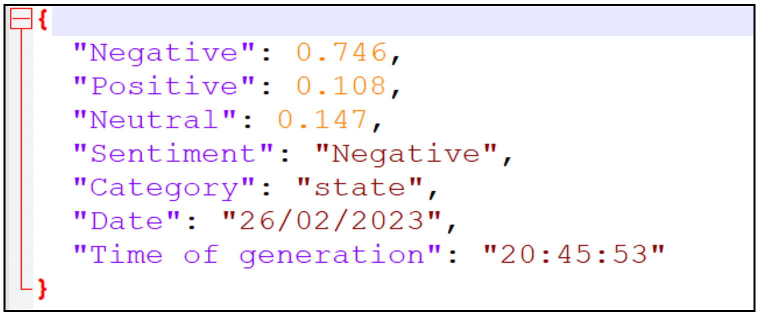


Monthwise News Trend Graph

The distribution of the range of predicted scores for each sentiment class is given in the scatter plots below:



A sample of the predicted output file is as follows:



```
{
    "Negative": 0.746,
    "Positive": 0.108,
    "Neutral": 0.147,
    "Sentiment": "Negative",
    "Category": "state",
    "Date": "26/02/2023",
    "Time of generation": "20:45:53"
}
```

REFERENCES

1. https://aclanthology.org/2021.icon-main.58/
2. https://aclanthology.org/2021.findings-emnlp.278/
3. প্রথম আলো | বাংলা নিউজ পেপার (prothomalo.com)
4. https://huggingface.co/csebuetnlp/banglabert
5. https://huggingface.co/sagorsarker/bangla-bert-base
6. https://www.kaggle.com/datasets/csoham/classification-bengali-news-articles-indicnlp
7. https://huggingface.co/sagorsarker/bangla-fasttext