Bangla News Article Headline Sentiment Analysis

Abstract

Automated Bangla article sentiment classification is basically finding out what type of sentiment does the data resembles with. Despite several comprehensive textual datasets are available for different languages, a few small datasets are curated on Bangla language. As a result, a few works address Bangla document sentiment classification problem, and due to the lack of enough training data, these approaches could not able to learn sophisticated supervised learning model. In this work, we created a large dataset of Bangla articles from **The Daliy Ittefaq**, which contains around 70k articles,but for sentiment ,we have labeled 1k headlines. This huge diverse dataset helps us to create a model by utilizing TF-IDF features, which finally predicts sentiments.

Methodology

We have used the tf-idf approach for sentiment analysis.

Dataset

For dataset, We have used the articles from famous Bengali newspaper “The Daily Ittefaq”.Almost 54K documents have been retrieved from the website of the newspaper for the years 2019 and 2020.Almost 9,00,000 sentences are present in the corpu ,but for sentiment ,we have labeled 1k headlines.

Summary

#Total Article count ==========> 53571

#unique unigram ==========> 252073

#Total Unigram ==========>13326446

#unique bigram ==========> 3718409

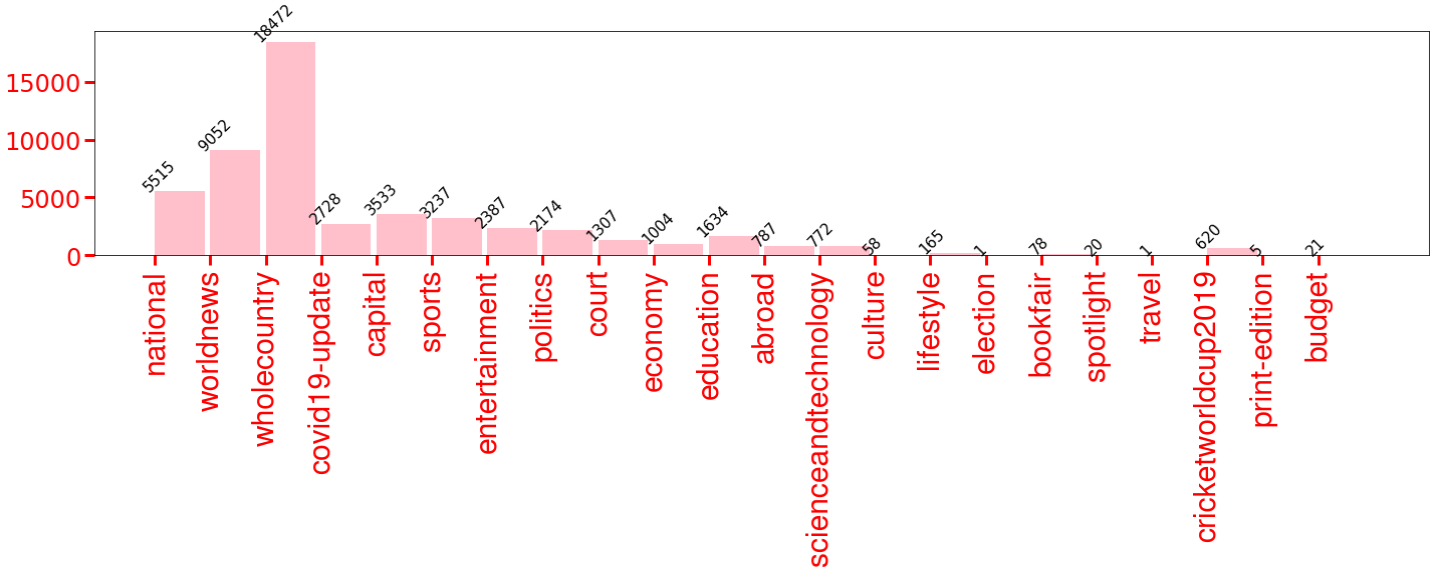
#Total Bigram ==========> 13579033

#unique trigram ==========> 8218523

#Total Trigram ==========> 24030358

Category Summary: **Total 22 Categories**

Type	News Count	Type (Bengali Version)
'national':	5515,	জাতীয়
'worldnews':	9052,	আন্তর্জাতিক
'wholecountry':	18472,	সারাদেশ
'covid19-update':	2728,	কোভিড-১৯
'capital':	3533,	রাজধানী
'sports':	3237,	খেলা
'entertainment':	2387,	বিনোদন
'politics':	2174,	রাজনীতি
'court':	1307,	কোর্ট
'economy':	1004,	অর্থনীতি
'education':	1634,	শিক্ষা
'abroad':	787,	বিদেশ
'Scienceandtechnology':	772,	বিজ্ঞান ও টেক
'culture':	58,	সংস্কৃতি
'lifestyle':	165,	জীবনযাপন
'bookfair':	78,	বইমেলা
'cricketworldcup2019':	620	ক্রিকেট বিশ্বকাপ -১৯



But for the sentiment analysis we could label 1k news headlines.

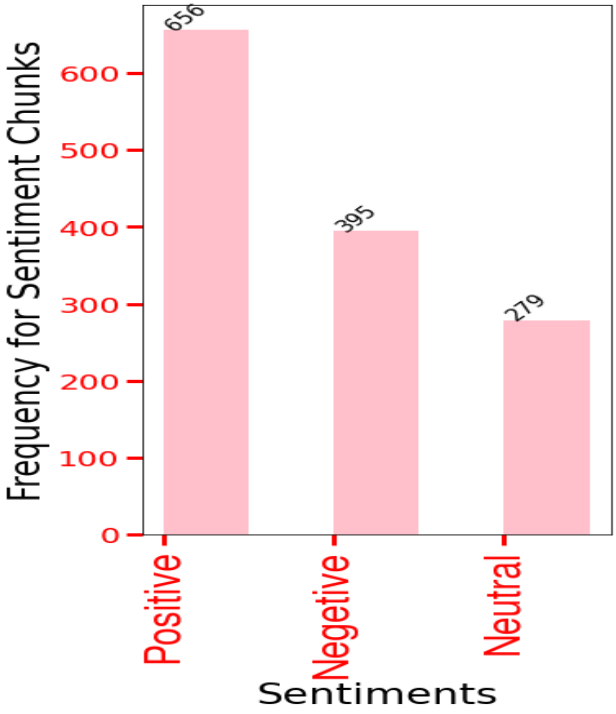


Fig: Sentiment dataset summary.

So, the methodology is consisted of following procedures:

A. Preprocessing

Preprocessing is very important in terms of raw data is concerned and especially in Bengali language where different punctuation marks are there.

a. Punctuation Removal

First of all, I have ensured the proper cleaning of each document by cleaning the punctuation marks. Again many Unicode were not detected, so they were needed to be cleaned as well.

b. Duplicate Sentence Removal

There were many duplicate sentences in the corpus as well. So I have made sure that they unique sentences will only be there clearing the duplicate ones.

c. Stop words Removal

There are many stop words in Bengali languages like "ও" , "এবং" ,"আর" etc.These stopwords are identified by computing frequency based unigrams.The most frequent words were the stop words so many stopwords were identified from the frequency based unigrams and from Bengali language point of view.

d .Stemming

There are several words in different formations like শহর, শহরে,শহরের etc. They basically denote the same meaning শহর but with the inclusion of terms like 's', 'es' it shows the different form. To remove redundancy, the conversion of several formations into an actual word is also needed.

e. Tokenization

Word tokenization is done here for each document to compute the tf-idfs of each word of each document which is described in the coming sections.

f.Small documents removal

Article of smaller lengths are removed in this case.

B. Bangla News Article Headline Sentiment Analysis

For sentiment analysis we have used the tf-idf approach.We have extracted the tf-idf of all headlines and used them for features for various machine learning models.

TF-IDF

TF-IDF is a measurement that symbolizes how important a word is to a document in a association of documents. Two metrics are multiplied in this regard. They are the term frequency and the inverse document frequency. Term frequency is how many times a word appears in a document divided by the length of the document in terms of word count and the inverse document frequency is the number of total documents divided by the number of documents that the word has appeared in.

$$\text{TF*IDF}= \frac{No\ of\ perticualr\ term\ in\ that\ doccument}{Total\ no\ nof\ term\ in\ that\ doccument} * \frac{Total\ No\ of\ doccument}{No\ of\ doccument\ where\ the\ perticular\ word\ appears}$$

…….(1)

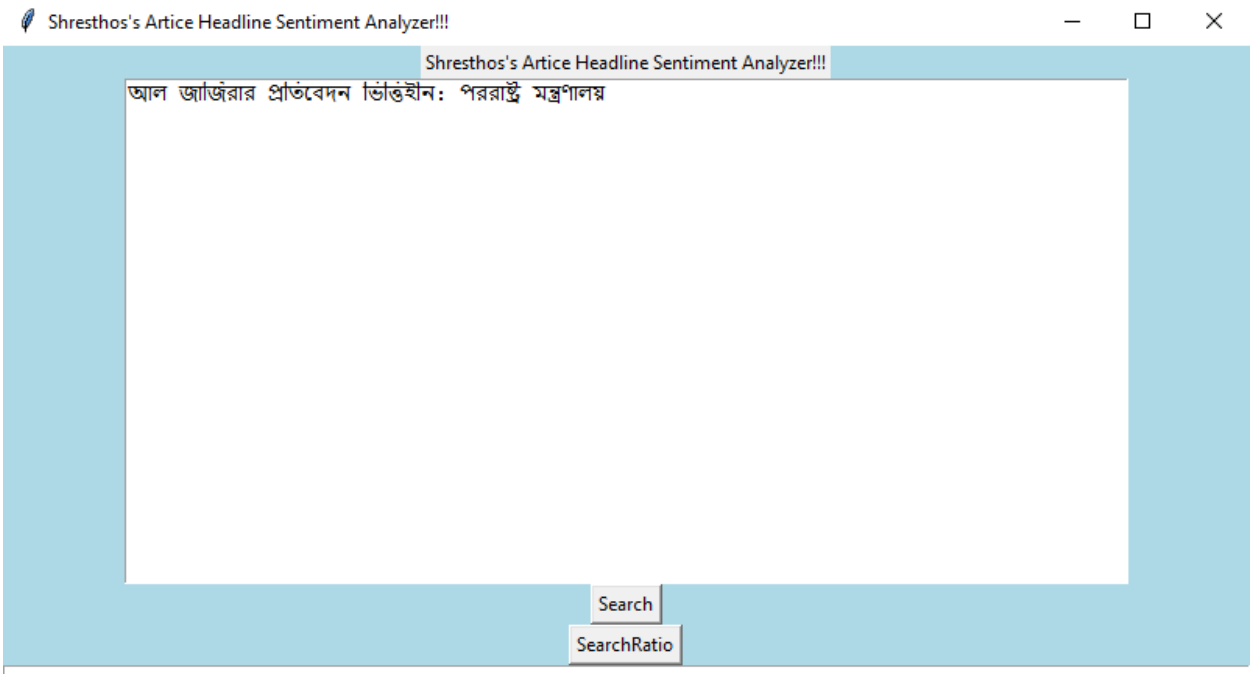
Fig: TF-IDF

C. Experimental result

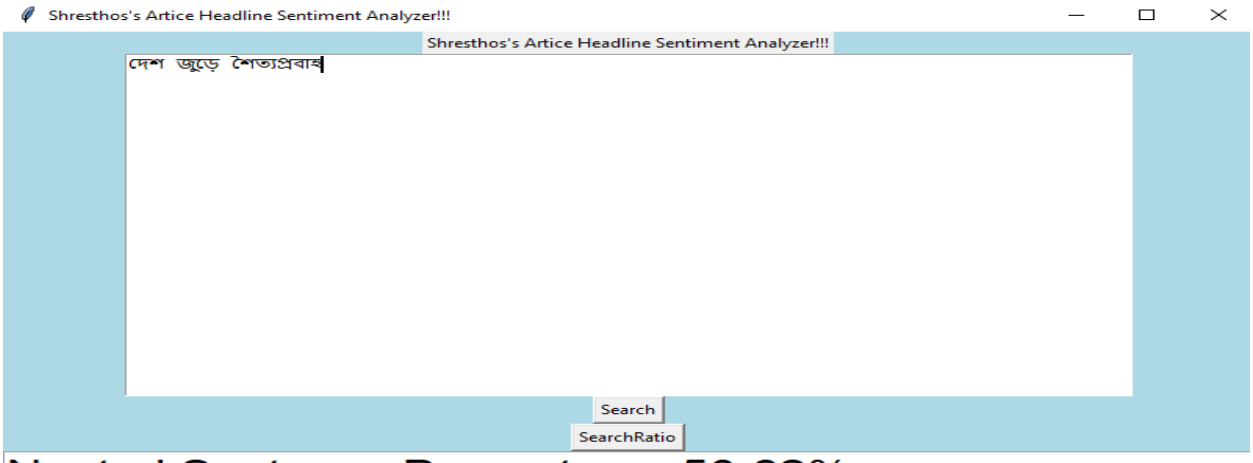
Accuracy of SVM: 82%

Accuracy of Nearesst neighbour:76%

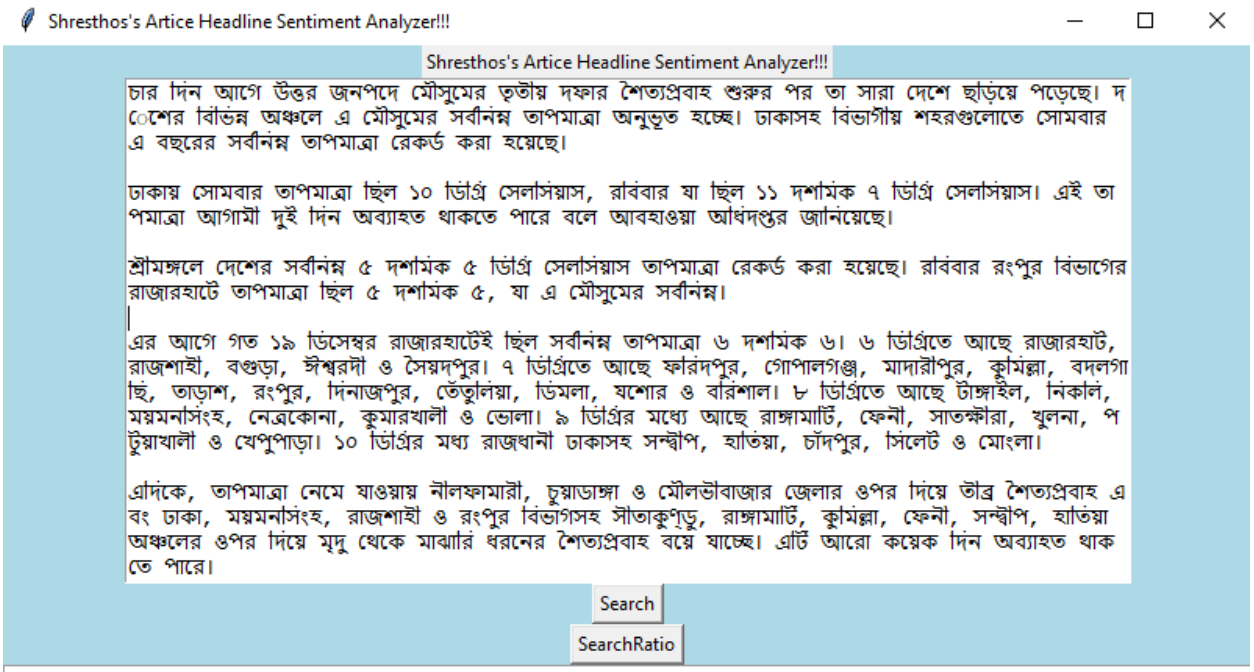
D.Application UI



Negetive Sentence Percentage :57.1%
Positive Sentence Percentage :32.95%
Neutral Sentence Percentage :9.95%



Neutral Sentence Percentage :50.62%
Negetive Sentence Percentage :45.51%
Positive Sentence Percentage :3.86%



Negetive Sentence Percentage :36.08%
Neutral Sentence Percentage :33.4%
Positive Sentence Percentage :30.52%