Proceedings of the SMART–2019, IEEE Conference ID: 46866
8th International Conference on System Modeling & Advancement in Research Trends, 22nd–23rd November, 2019
College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India

Performance Measurement of Multiple Supervised Learning Algorithms for Bengali News Headline Sentiment Classification

Md. Majedul Islam[1], Abu Kaisar Mohammad Masum[2], Md Golam Rabbani[3], Raihana Zannat[4] and Mushfiqur Rahman[5]

[1,2,3,5]*Dept. of CSE, Daffodil International University, Dhaka, Bangladesh*
[4]*Dept. of Software Engineering, Daffodil International University, Dhaka, Bangladesh*
E-mail: [1]majedul15-6784@diu.edu.bd, [2]mohammad15-6759@diu.edu.bd, [3]golam15-204@diu.edu.bd, [4]zannat.swe@diu.edu.bd, [5]mushfiqur.cse@diu.edu.bd

Abstract— The reading newspaper is a common habit in today's life. Before reading news article all are focused on the news headline. Understanding the meaning of news headline everybody can easily identify the news types. That means the containing news article provides positive or negative news. Analysis of the sentiment of the news headline is a good solution for this kind of problem. Sentiment Analysis is a chief part of Natural Language Processing. It mines any kinds of opinion and set the sentiment of any text. We proposed a method for Bengali news headline sentiment measurement with different kinds of the supervised learning algorithm and their performance. Firstly, we set sentiment of each news headline then used the classification method to predicting the news headline which was containing a positive or negative headline. After all, Bengali is one of the most used languages in this world. A lot of research work done previously in a different language but very few in the Bengali language. So, increasing the Bengali language research resource need to develop different kinds of tools and technology.

Keywords: Sentiment Analysis, Natural Language Processing, Opinion Mining, Bengali News Headline Sentiment

I. Introduction

Any human language problems are solved by NLP in AI research fields. It grasps the concept of human language problems and tries to provide a solution for the machine. The machine learning algorithm is the most usable algorithm for understanding the NLP problem with the solution. Machine learning is a concept which meaning an automatic learning system. A few approaches have in machine learning such as supervised, unsupervised and semi-supervised learning. In supervised learning provided labelled data with input and output but in unsupervised learning provided only unlabeled input data and out will generate from input data. Semi-supervised learning is made from both combinations where the label and unlabeled data have in mixed.

Peoples express their opinion after reading any kinds of text and given the opinion will be negative, positive or neutral. Sentiment analysis helps to appreciate the opinion of providing text documents. News headline is a short text which contains the gist of the news. Everybody follows the headline before reading the news, at that time they understand the sentiment of news. In this paper, we introduce a method for Bengali news headline sentiment analysis using the multiple machine learning algorithms. We determine the news headline sentiment by 0 and 1 where 0 consists of negative news and 1 are positive news. After preparing the data, trained by multiple supervised learning classification algorithms which provide a predicted output with good accuracy.

II. Related Work

Sentiment analysis is the most usable research in natural language processing. Formerly various research work has done successfully in this field. This section we have discussed some related work which helps us to complete our research purpose.

A. News and Blogs Sentiment

News sentiment analysis is different from normal text sentiment analysis such as a review analysis, Balahur A et al [6]. The terminology of the news article apparently does by the writer. In review analysis, the related word is figured but in news, it's difficult to find out for large and complex description. Make a short and long word essence to find out positive and negative news sentiment. Godbole et al. [2] attach a scoring rate to express the positive either negative news and blogs sentiment offers a solution for large text substance. Analyzing this sentiment will help to indicate the future acclaim and advertise of news and blogs. Fu Y et al. [5] proposed a methodology for travel news sentiment analysis. They analyze the key factor for china tourism and provide better predictive accuracy for future tourism research study.

B. ML Algorithms for Sentiment Analysis

ML approaches provide a satisfactory result and accuracy for review sentiment. Naive Bayes and SVM give the best performance from other algorithms, Jagdale

et al [7]. Twitter is the most important source of sentiment analysis for social media. Here opinion is divided into three categories such as happy unhappy and neutral. Kurnaz et al. [8] proposed a system with Sparse Autoencoder algorithm which gives 0.98 accuracies for twitter data sentiment analysis. For sentence-level news text, SVM and Naive Bayes give 96.46 % and 94.16% accuracy, Shirsat et al. [1].

Work with Bengali text in any NLP research area is challenging. Data processing and preparation is different from other languages. This paper we try to apply different approaches of ML to provide an accurate future news headline sentiment prediction. Where different algorithm provides different accuracy with a correct prediction result.

III. Methodology

Machine learning approaches help to solve NLP problems. In natural language processing important problem such as text analysis, sentiment analysis, speech to text conversion, text summarization, image to text conversion, language to language translation all is solved using machine learning technique. Sentiment analysis is also an important part of natural language. Mine the opinion from the text document is the main concept of solving the sentiment analysis problem. This research work we follow the NLP and ML approaches to solve the Bengali news sentiment classification. Given below a workflow for this research work.

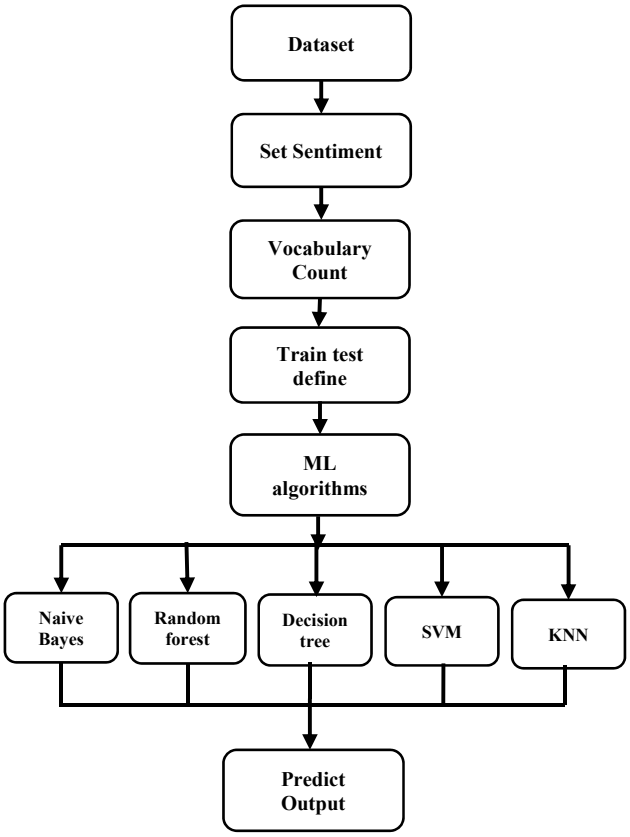

Fig. 1: Working Flow for Bengali News Headline Sentiment

A. Data Collection and Dataset Properties

Newspaper headline estimation expectation is the primary centre point in our research work. So a marked dataset is required for the conclusion characterization. We gather information from Bengali paper "prothom alo" utilizing web scratching system with python scripting. After collecting the data we set the sentiment of the headline. Headline sentiment divides two types where 0 means negative headline and 1 means positive headline. Dataset properties resemble given below.

a. Total data 1619
b. 11 types of news
c. 1109 positive headline and 510 negative headline
d. Minimum & maximum word length 1 and 14.

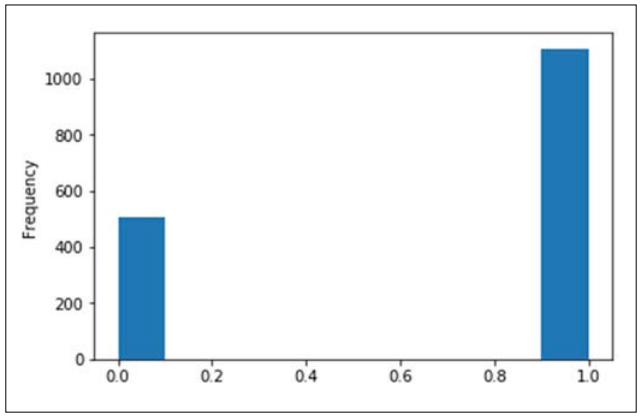

Fig. 2: Frequency of Positive & Negative Sentiment

In figure 2, x-axis contains the frequency of the negative and positive news headline where y-axis contains the positive and negative news sentiment.

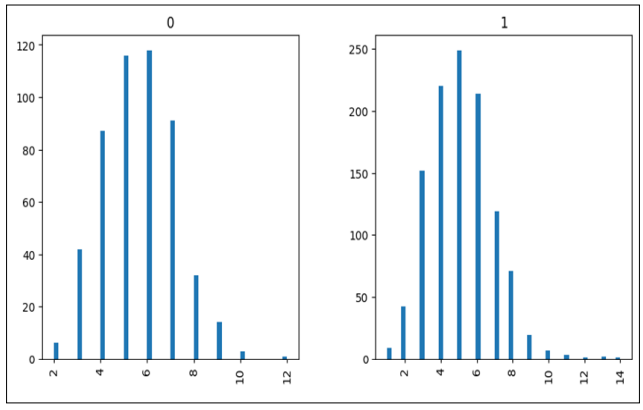

Fig. 3: Word Length of Positive and Negative Headline

In figure 3, x-axis contains the number of headlines and y-axis contains total length. The maximum length of negative news headline is 12 and the amount of headlines is 2. Minimum length of negative news headline is 2 and the number of headlines is 3. For positive news maximum number of headlines, text length is 14 and the total number of headline 12 where the minimum number of text length is 1 and the amount of headlines is 5.

B. Data Preprocessing

The procedure of Bengali content information is troublesome from the procedure of different dialects information. The machine couldn't recognize Bengali language characters or images naturally. To evacuate an undesirable character, space letter or digit, Bengali accentuation needs to characterize Bengali Unicode of the characters. The scope of Bengali Character Unicode is 0980-09FF. Another part of preprocessing is needed to expel space from the line and evacuate the stop words. For stop words remove we collect all Bengali stop words and save into a file then remove stop word from the dataset.

1) Add Contractions

Using a short form of a word is knows as contraction. There are a few contractions in the Bengali language. Such as, "ডা." is the short form of "ডাক্তার". Before pre-processing all of this contraction was added to the dataset text.

2) Stop Word Remove

In preprocessing removing stop word is very important. Stop word contains the most common word in a text or document. So in natural language processing stop, words are removed from the text for any language modelling. There are many stop words in the Bengali language such as আছে, আমরা, এখন etc.

3) Unwanted Character Remove

A machine can't understand a rare character or word. So in the pre-processing step remove unwanted characters is very important. In Bengali text whitespace, punctuation, some digits are included in unwanted characters.

C. Vocabulary Count

For vocabulary count, we use Count Vectorizer. It counts the split word which is showing up in dataset. Then uses the weight in input for vocabulary count. After the count, we fit and transform input with vocabulary.

D. Train Test Data

After ensuring the fit of the input parameter dataset needs to train for machine learning. Supervised learning way is required for classification technique. Because in the dataset label and input-output given. Then define test dataset to remove the unbiased assessment. In the model train, almost 85%data was given and for test dataset, 15% data with 101 random state are defined.

E. Machine Learning Algorithms

Supervised learning algorithms are used to solve all classification problem. The classification problems are following true and false logic. If the predicted input is positive it's true otherwise it's false. All of the predicted output is depending on the input label. Suppose x is an input variable and y is an output variable. So, output variable y is dependent on the input variable x. The classification function f will be,

$$y = f(x) \tag{1}$$

Headline sentiment is a classification problem. Input news headline text identifies the output. The output contains sentiment of the news. Classification algorithm helps the true prediction of the output result. For the experiment, we used five classification algorithms with a suitable parameter. Briefly discussed in below about uses algorithms.

1) Naive Bayes Classifier

This algorithm is used to calculate the probability of the classification problem. In our research, we use multinomial NB which is a distinct classifier used for multinomial disposal. Suppose the probability of the input feature is,

$$p = (x_i|c_j) \tag{2}$$

Here, x is the independent variable and c is class.

2) Random Forest Classifier

Random forest classifier depends on decision tree logic. In each classification prediction, everyone works a separate decision tree. The maximum number of the tree for class value is predicted output for this classifier. Average of single decision tree make a random forest classifier. So the equation for this classifier will be,

$$f_{i_i} = \frac{\sum_{j \in total\ tree} norm f_{i_{ij}}}{T} \tag{3}$$

Here, f_{i_i} = important factor from all tree
norm $f_{i_{ij}}$= normalize factor from tree
T= tree number

3) Decision Tree Classifier

The decision tree is the most capable and usable classification algorithm. Output generated by yes and no technique basis. All value depends on the input label then generated the prediction.

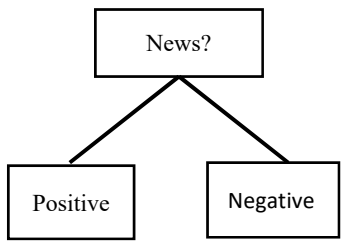

Fig. 4: Decision Tree for News Sentiment

IV. Nearest Neighbors Classifier

KNN is a non-parametric approach for classification algorithms. Output value calculated by the value of k which means the nearest value of k. where k is a parameter for find related output. k search the closest values for the providing

parameter from the dataset. In our experiment, we use the value k=3 and provide a good result. Each instant is selected by the distance measurement. If the instance distance is near to the k value is put in the nearest neighbours then calculate the minimum distance from the value which will be the final value

4) Support Vector Machine Classifier

Support vector machine is the most useful method for sentiment analysis classification. Because it provides the best accuracy for this type of problem. The hyperplane is used in each support vector machine classifier. Each hyperplane divided each dataset into two-part. The hyperplane is worked based on the kernel where the kernel represents some algebraic calculation. We use SVC kernel for our classification problem. SVC contain a vector classifier.

F. Model Discussion

Machine Learning algorithm provides a better result for sentiment analysis problem. We have seen all previous research that Support Vector Machine and Naive Bayes algorithm provide accurate result rather than other supervised learning algorithms to classify any sentiment analysis problems. In this research, we try to find out the best algorithms for Bengali news headline sentiment classification based on some supervised learning algorithm. And finally, selected the algorithms for classifying the Bengali news type depend on algorithm prediction.

The necessary steps of the model are given below for choosing the classification algorithm.

Step 1: Read the news headline dataset.

Step 2: Set the news sentiment, negative news = 0 and positive news = 1.

Step 3: Pre-process the headline text.

Step 4: Count the vocabulary for using as model input.

Step 5: Fit and Transform the vocabulary.

Step 6: Divide the train and test.

Step 7: Define the machine learning algorithm and train the model.

Step 8: Check the algorithm accuracy and prediction result. If the prediction of the algorithm is equal to the actual prediction result then select the algorithm for headline classification.

All of these steps are following for news headline classification based on the using algorithms.

V. Experiment and Output

This experiment, after dividing the test and train dataset we applied multiple machine learning algorithms. Using approaches are Naive Bayes, SVM, Random forest, Decision tree, and K-nearest neighbours. Previous all experiment in sentiment analysis Naive Bayes and SVM contribute the best accuracy. Similarly in this experiment, 75% accuracy from SVM and 73% from Naive Bayes classification algorithm which is the best from the other three algorithms. Random forest commit 69%, KNN commits 68% and Decision tree commits 60% accuracy for positive and negative news classification. In table1 discuss the performance and accuracy for the algorithms.

Table1: Performance for Bengali Headline Sentiment Analysis

Approach	Sentiment	Precision	Recall	F1-score	Accuracy
Naive	0	0.55	0.24	0.34	73%
Bayes	1	0.75	0.92	0.83	
SVM	0	0.68	0.21	0.33	75%
	1	0.75	0.96	0.84	
Random	0	0.44	0.36	0.39	69%
Forest	1	0.76	0.82	0.79	
Decision	0	0.33	0.40	0.36	60%
Tree	1	0.73	0.67	0.70	
KNN	0	0.45	0.39	0.41	68%
	1	0.76	0.79	0.78	

In figure 5 the bar chart displays the accuracy comparison for applying algorithms.

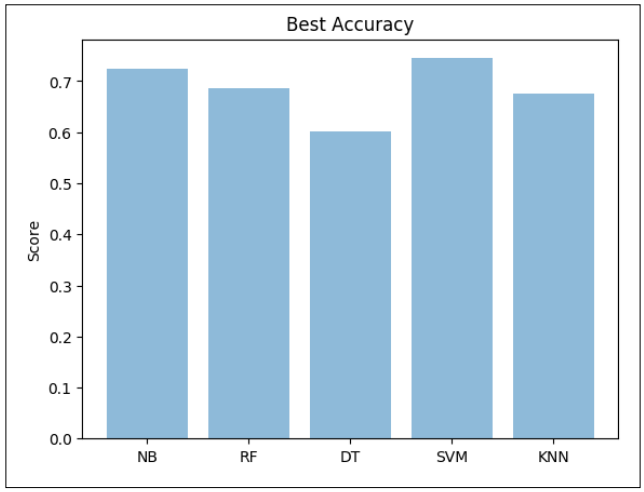

Fig. 5: Accuracy Chart for ML Algorithms

Now we have used another table to check the classification result with a Bangla News headline. Where all of that applied algorithm predicts the accurate output.

Headline = "রাজবাড়ীতে মোটরসাইকেলে দুর্ঘটনায় কলেজে ছাত্ররে মৃত্যু" in English ("College student dies in a motorcycle accident in Rajbari")

Actual Prediction = 0

News Type = Negative News

TABLE 2: NEWS CLASSIFICATION FOR THE GIVEN HEADLINE

Prediction	Sentiment	News Type	News Classification
SVM Prediction	0	Negative News	Correct
NB Prediction	0	Negative News	Correct
DT Prediction	0	Negative News	Correct
RF Prediction	1	Positive News	Incorrect
KNN Prediction	1	Positive News	Incorrect

Table 2 shows the classification result for the given headline. The provided headline is negative news and the predicted value is 0. So, if actual output is equal to the predicted output then that algorithm choose for news headline classification. Here SVM, Naive Bayes and Decision Tree provide actual prediction others two give the wrong prediction. But others sample only SVM and Naive Bayes provide an accurate prediction. Finally, SVM and Naive Bayes classifier are used for Bengali news headline sentiment classification.

VI. CONCLUSION AND FUTURE WORK

This experiment work proposed a methodology for making a Bengali news feature conclusion analyzer utilizing numerous ML Algorithms. Since no machine gives a precise outcome notwithstanding yet utilizing calculations gives some exact outcome. Utilizing the proposed technique have effectively Identify the positive and negative news for Bengali newspaper. The precision of applying need to build which is in our future work. There are two or three imperfections in the proposed system. One is less dataset. For accurate result need a large dataset but manually sentiment provide is a lengthy process. The vocabulary of the dataset is low so for achieving a good accuracy need to increase vocabulary. Machine learning algorithm shows good performance for Bengali data but not better but in the English language, the problem gives it's better performance. So in future, there is the workspace to improve accuracy for Bangla text with the excellent outcome from ML algorithms.

ACKNOWLEDGEMENT

We acknowledge and thanks to our DIU NLP and Machine Learning Research Lab for their total assist. Special thanks for our Computer Science and Engineering department for help to complete the work and provide the facility for research.

REFERENCES

[1] Shirsat, Vishal S., Rajkumar S. Jagdale, and Sachin N. Deshmukh. "Sentence Level Sentiment Identification and Calculation from News Articles Using Machine Learning Techniques." In Computing, Communication and Signal Processing, pp. 371-376. Springer, Singapore, 2019.

[2] Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs." Icwsm7, no. 21 (2007): 219-222.

[3] Shapiro, Adam Hale, Moritz Sudhof, and Daniel Wilson. "Measuring news sentiment." Federal Reserve Bank of San Francisco, 2018.

[4] Zhang, Wenbin, and Steven Skiena. "Trading strategies to exploit blog and news sentiment." In Fourth international aAAI conference on weblogs and social media. 2010.

[5] Fu Y, Hao JX, Li X, Hsu CH. Predictive Accuracy of Sentiment Analytics for Tourism: A Metalearning Perspective on Chinese Travel News. Journal of Travel Research. 2019 Apr;58(4):666-79.

[6] Balahur A, Steinberger R, Kabadjov M, Zavarella V, Van Der Goot E, Halkia M, Pouliquen B, Belyaeva J. Sentiment analysis in the news. arXiv preprint arXiv:1309.6202. 2013 Sep 24.

[7] Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh. "Sentiment analysis on product reviews using machine learning techniques." In Cognitive Informatics and Soft Computing, pp. 639-647. Springer, Singapore, 2019.

[8] Kurnaz, Asst Prof Dr Sefer, and Mustafa Ahmed Mahmood. "Sentiment Analysis in Data of Twitter using Machine Learning Algorithms." (2019).

[9] Chowdhury, SM Mazharul Hoque, Priyanka Ghosh, Sheikh Abujar, Most Arina Afrin, and Syed Akhter Hossain. "Sentiment Analysis of Tweet Data: The Study of Sentimental State of Human from Tweet Text." In Emerging Technologies in Data Mining and Information Security, pp. 3-14. Springer, Singapore, 2019.

