Technically Co-Sponsored by IEEE UP SECTION (INDIA)

NURTURE Education Solutions

Marwah Infotech — Website Development & Software Solutions — Augur — SAP Partner

CETPA — Because Knowledge Matters — ISO 9001:2015 Certified

TMU Moradabad

Proceedings of the
2019 8th International Conference on

System Modeling & Advancement in Research Trends

(22nd–23rd November, 2019)

SMART–2019

(IEEE Conference ID: 46866)

Technically Sponsored by IEEE Uttar Pradesh Section

Editor-in-Chief

Prof. Rakesh Kumar Dwivedi
Professor & Principal,
College of Computing Sciences & Information Technology,
Teerthanker Mahaveer University, Moradabad, UP, India

Editors

Dr. Ashendra Kr. Saxena
Associate Professor & HOD
CCSIT, Teerthanker Mahaveer University,
Moradabad, Uttar Pradesh, India

Dr. Danish Ather
Associate Professor
CCSIT, Teerthanker Mahaveer University,
Moradabad, Uttar Pradesh, India

Dr. Danila Parygin
Associate Professor
Volgograd State Technical University
Russia

Dr. Vibash Yadav
Associate Professor,
Rajkiya Engineering College, Banda
Uttar Pradesh, India

Organized by

College of Computing Sciences & Information Technology

TEERTHANKER MAHAVEER UNIVERSITY
Moradabad, India

www.smartconference.co.in, www.tmu.ac.in, smartconf@tmu.ac.in

Acknowledgment

Technically Co-Sponsored by

- IEEE UP Section

Sponsored by

- iNurture Education Solutions Private Limited, Bangalore, India
- CETPA Pvt. Ltd.
- Teerthanker Mahaveer University

ISBN: 978-1-7281-3243-3

DISCLAIMER

The opinion expressed and figures provided in this proceedings of SMART–2019 are the sole responsibility of the authors. The organizers and the editor bear no responsibility in this regard. Any and all such liabilities are disclaimed.

Published by

Prof. Rakesh Kumar Dwivedi, General Chair, SMART–2019 and Principal College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India. Tel.: +91-9837771280, E-mail: smartconf@tmu.ac.in

Publishing Consultancy

EXCEL INDIA PUBLISHERS

91 A, Ground Floor
Pratik Market, Munirka, New Delhi–110 067
Tel: +91-11-2671 1755/ 2755/ 3755/ 5755
Cell: 9899127755, 9999609755, 9910757755
Fax: +91-11-2671 6755
E-mail: publishing@groupexcelindia.com
Web: www.groupexcelindia.com

Typeset by
Excel Prepress Services, New Delhi–110 067
E-mail: production@groupexcelindia.com

Printed by
Excel Printing Universe, New Delhi–110 067
E-mail: printing@groupexcelindia.com

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.

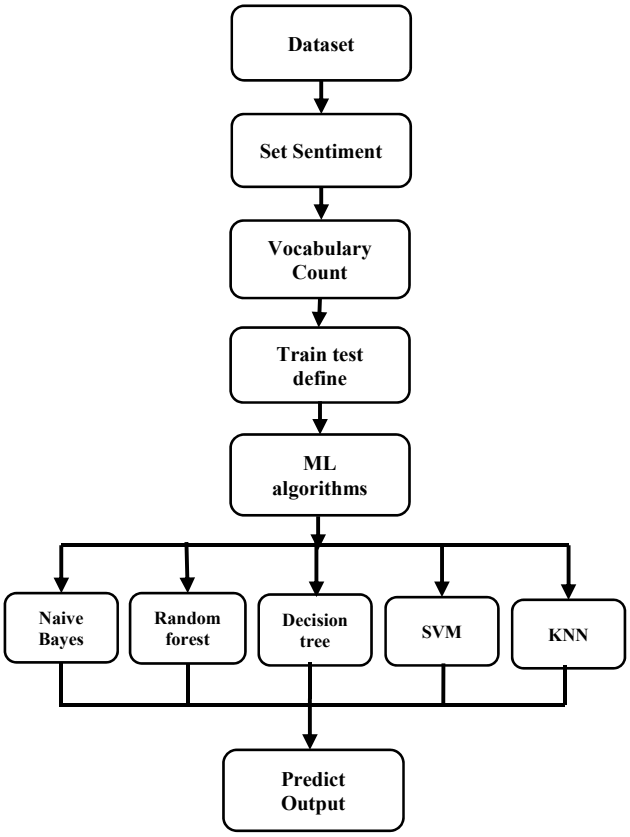

Fig. 1: Working Flow for Bengali News Headline Sentiment

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.

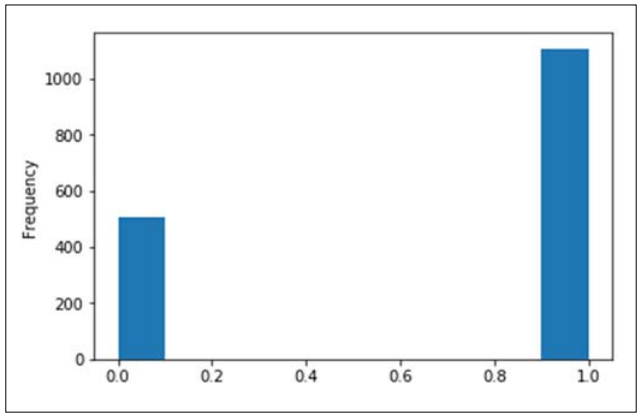

Fig. 2: Frequency of Positive & Negative Sentiment

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.

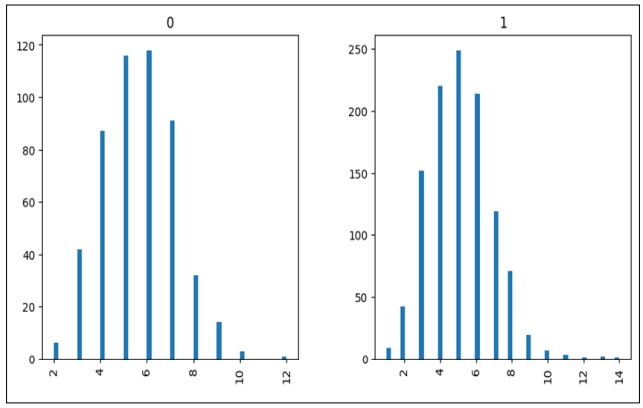

Fig. 3: Word Length of Positive and Negative Headline

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.

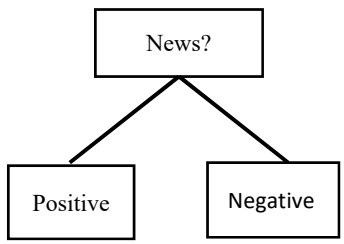

Fig. 4: Decision Tree for News Sentiment

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.

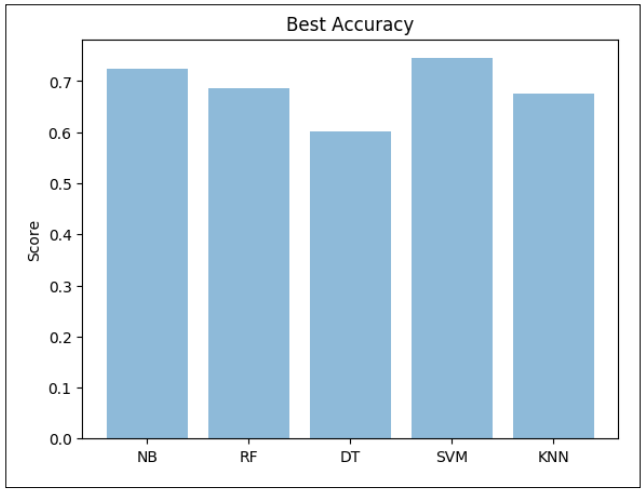

Fig. 5: Accuracy Chart for ML Algorithms

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.

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

Comparative Sentiment Analysis using Difference Types of Machine Learning Algorithm

Rakib Hossain[1], Fowjael Ahamed[2], Raihana Zannat[3] and Md. Golam Rabbani[4]
[1,2,4]*Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh*
[3]*Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh*
E-mail: [1]rakib15-6802@diu.edu.bd, [2]fowjael15-7045@diu.edu.bd, [3]zannat.swe@diu.edu.bd, [4]golam15-204@diu.edu.bd

***Abstract*—In today's world business are becoming online based. Companies sell their products and seek for consumer's feedback. When all the consumer writes their review about that's a product, It's becomes difficult to say that product is good or not based on their review. That's where Deep learning come. By using this, we can extract opinion or sentiment from the text which is written by the consumer. This is sentiment analysis. It can classify the emotional status of that review. Our project detects opinion from consumer's review whether it is good or bad. We use SVM, Naive Bayes algorithm and some methods. We use the Naive Bayes algorithm because we want to know how often words occur in the document. And then we use SVM for classifying whether words are positive or negative. For our researching purpose, we use the Amazon consumer review data set, which was available online. Some methods that we are using for preprocessing and cleaned the document where just words are left. We trained our model so well with twenty-four thousand data. So, it will give us the best accuracy and we make this model with the best algorithm and after that, it gives the accuracy of 98.39%. This project will help us in real life when we are having trouble with product reviews. Our machine will help us to determine which review is good and which review is bad and make a category of a positive and negative review and saves our time.**

Keywords: NaïveBayes, SVM, KNN, Polarity, Sentiment, Positive, Negative, Word, Paragraph, Accuracy

I. Introduction

Technology is the most important thing in today's world. Every human depend on this more and more. So the huge amount of data are created every moment. This data are used to develop the product by pre-processing. Opinions are the most important data for improved the product and it is the big research data for the recent world. It is very important to classify them like negative or positive. As an example "It's not suitable for him", it's must be negative type opinions. But the positive type of word 'suitable' will give a positivity. But the word 'not', are full changes the meaning of the sentence.so, it proves that we can't decide a sentence is a positive or negative basis on some keywords. It is important that a word is used after or before the keyword must be the consideration when made meaning of the sentence.

It is very elaborate type problem to divine the sentiments given the text, but this problem solved using by Naive Bayes classifier, K neighbors classifiers and support vector machines. In this paper, I represent some classifier methodology like SVM and Sequential model, feature selection, words emphasizing and effective negation handling which is improved the accuracy of the result in sentiment analysis.

II. Related Work

Thousand of data analyzer all over the world are trying to pursuit to increase the accuracy of sentiment analysis. Paul Ferguson team pursuit analysis to improve the accuracy of sentiment analysis based on paragraph level in 2009 [6]. In 2009, Yelena Mejova working with emotionally-charged text based on sentiment. Also working this field to develop sentiment accuracy, Christos Livas, Konstantina Delli and Nikolaos Pandis Invisalign patient testimonials on YouTube as well as the sentiment of the comments. Francesc Alías and Alexandre Trilla working with speech based sentiment in 2013 [7]. Aspect-based review analysis was done by Devina Ekawati and Masayu Leylia Khodra in 2017 [8].

Nurulhuda, Zainuddin and his team [1] worked sentiment analyzing using SVM. They worked using benchmark dataset to train the classifiers. They used to different weighting scheme and N-grams to extract the classical feature. They using square feature selection for improving the classification accuracy.

Pang, B., & Lee, L [2] represent a model for classifying the movie reviews. They used machine learning to find out the difference between polarity classification and subjectivity detection and proposed a method for text-categorization. Show that the Naïve Bayes algorithm is more effective to show the result using subjectivity detection to shorter reviews.

Abbasi, A., Chen, H., & Salem, A. [3] proposed a model for classifying multiple language web forum reviews. To improve the performance the author used entropy weighted algorithm. They used SVM to get the to get higher performance with feature selection methods with high accuracy more than 90%.

Zhuang, L., Jing, F., & Zhu [4] proposed a method produce some feature automatically in a movie review analysis system using a multi knowledge-based approach.

They combined the wordnet, statical analysis, and knowledge of movies. Their model was so effective it's proved that final output.

S.M. Mazharul Hoque Chowdhury and his team [5] approach a method to analyze a text in the paragraph level. The implementation of this method using a bag of words and priority based on the lexical analysis.

So, we proposed a method which is given better accuracy other than analysis.

Nadeem Akhtar and his team [9] are reviewing the hotel feedback and providing information that may miss scores. The comments and metadata were crawled from the website and grouped according to some of the common aspects into predefined categories. Then the subject modeling technique (LDA) is applied to define hidden information and features, accompanied by an evaluation of feelings on confidential phrases and summarization.

Omar Raghib, Eshita Sharma, Tameem Ahmad and Faisal Alam [10] discuss a sequence of steps to be taken in this paper to evaluate the speech signal to understand their feelings, outlining some of the best techniques available for each stage at present. Throughout transmission, disturbances such as background noises exacerbate and difficult speech recognition. This paper offers a framework for conducting speech recognition of different emotions.

Vanshika Varshney, Aman Varshney and their team [11] are exploring the techniques of different machine learning algorithms to deduce a user's personality from their social media activities. Using three algorithms to compare the results of these three algorithms, namely SVM, KNN and MNB. Eventually, the author provides all three algorithms with a cumulative performance.

Palak Bansal, Somya, Nazar Kamaal, Shreya Govil, Tameem Ahmad [12] research is an effort to summarize the product reviews of consumers in a more usable and concise version that can help other users make their decisions. Web reviews are crawled of product, the first detection of product features will be performed every time after extraction and therefore polarity will be identified, i.e. either a review is positive review or a negative review. The description of all product features will be produced after the calculations.

Istuti Singh, Anil Kumar Sahu [13] worked on this article, a study to examine the behavior of stone columns used in various building types, such as oil storage tanks, embankments, houses, etc. The effect of stone columns without encased and enclosed on several construction forms is being examined. Also checked the effect of different diameters with different depths in the soil. Various types of geosynthetics are used for the enclosure to enhance the performance. A variety of mathematical and physical methods were performed to predict the settlement of foundations reinforced with stone base. In this article, there are also several theories from past to present that help in understanding stone column enhancements to improve soft soils. Physical simulation has a major role in the development of geotechnical properties.

III. Some Issues in Sentiment Analysis

It is a very popular topic nowadays. When it's research using machine learning, it's faced some issues analyzing the text. An example an opinion gives a positive review in one nation but it's different from the other nation. Here, some common issues in sentiment analysis:

- Interrogative sentence gives us incorrect score because it's didn't identify properly which is a positive or negative opinion.
- Some opinions are didn't identified. Because it's not used any sentimental word.
- Some opinions can be spam when analyzing reviews and it provides negative score although it's a positive review.
- Sarcastic opinions are a challenging part of the analysis to be handled. Some of the sarcastic opinions give us negative meaning but it looks like a positive review. That's the case it provides a different result. Although a maximum number of the sarcastic sentence are positive.
- Different types of language give different meaning in a single word. It's made hesitation gives accurate output.
- Different types of language use in a single language, it's doesn't healthy for accurate accuracy.

IV. Proposed Model

A model is proposed when offered and an argument the element of the data set model and also discussed. We have

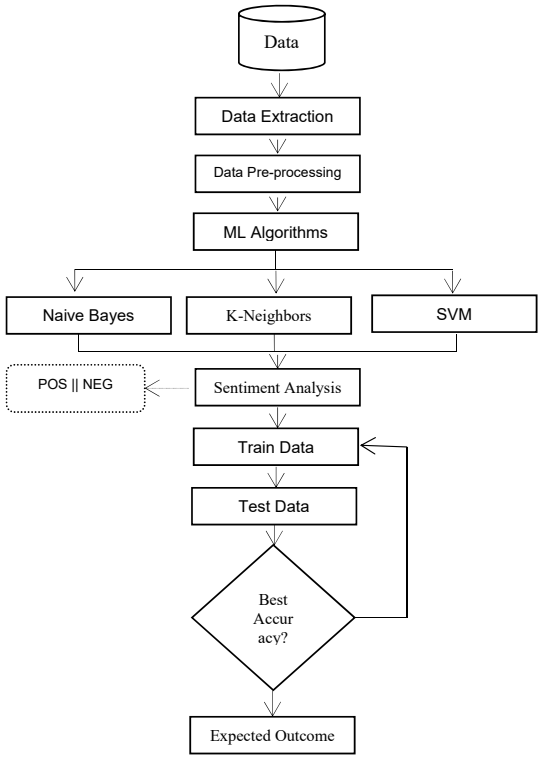

Fig. 1: Proposed Model

proposed a model which get the input in text set. This text set has some opinion as review or comment which used in the analysis for taking sentiment. Two major layers in our proposed model one are data processing and another one is analyzing sentiment. Data processing layer discuss the collection of data, pre-processing data and mad prefer for sentiment analysis. Another layer processes the data for the train and testing in a method and gets the expected output. We will be discussed it's briefly in the following section.

V. Methods Description

It's an important part in this research where we applied some method before data training and testing called data preprocessing.

A. Text Tokenization

Text tokenization is a way to convert a sentence into the separate section. In this methodology, we convert a sentence into words where different types of punctuation are included. Every section is separate consider the white space in a sentence. When it's getting a full stop then consider the line reaches the endpoint. An example: "If you want something new, you have to stop doing something old.". In this example tokenize the whole sentence like 'If',' you',' want',' something',' new','(,)', 'you',' have',' to',' stop',' doing',' something',' old','(.)' where ',' and '.' Are punctuation and '.' The dot indicates the endpoint of this line. These whole things we complete using this NLTK method which tokenizes the many languages with English. In NLTK, Takes a string in a text and perform the following this task like remove all punctuation, remove all stop words and finally return the clean text into list words.

B. Word Filtering

After tokenization, unexpected word, number, and punctuation are removed that will not be effective in classification. So, at first, removed the punctuation and numbers which are not to require in classification. After remove that finally unwanted word are removed which didn't detect the positivity or negativity in a sentence. If we consider the previous example then we removed the all unauthorized object and get the filtering word are ' want', ' stop', ' doing', ' old', ' new', ' something'. NLTK is performing the following word filtering task.

VI. Classification

Proposed a method for sentiment analysis before we need to a data set which used for analysis.so, Takes in a string of text input for analyzing which are collaborating a sentence or paragraphs. The text or reviews are loaded into the system before pre-processing. When pre-process the data set that's time to remove all the white space with punctuation, because it's not needed us for the sentiment. We also remove all the stop words before splitting the sentence into word, after that all the words are converted into lower case context. Finally, all the data return for text processing.

Before text processing, need to separate the positive word, negative word and also objective. Sentiment analysis phase identification the available sentiment word. After identification and separation word we should make the label of data set, positive is one (1) and negative is zero (0). Next, we vectorize our input variable using count vectorizer function which returns a vector array. Finally modified the data set compressed spares row format using transform function ().

A. Naïve Bayes Classifier

Naive Bayes classifiers are the collaboration of multiple classifiers algorithm based on Bayes theorem. It is not a single algorithm but it is a collection of sub familiar algorithm where share the common principle. The naïve Bayes algorithm is probabilistic classifiers for classification. It's all feature are independence assumption if compare between predictor. This model is no complicated iterative parameter estimation that's why it's very useful for every big dataset and it is easy to build a model.

Posterior probability P(t|z) is calculating by using Bayes theorem from P(t),P(z) and P(z|t), where P(t) represents the prior probability of class, P(z) is prior probability of predictor and P(z|t) provides the probability the text appears in this class. This forwardness is called class conditional independence.

Consider the following equation is:

$$P(t|z) = \frac{P(t)P(z|t)}{P(z)} \quad (1)$$

Where,

Class t provides predictor z for posterior probability P(t|z). The value of class t is identified as Positive and Negative where z is a sentence. The value of z is called true when the probability of t is true. because P(t) represents the prior probability of class. P(wi| z) is the probability of the i th feature in given class t where text appears z forget the value of t to maximize P(t|z). We need to train the parameter P(t) and P(wi|t).

B. SVM Classifier

There are so many different types of the algorithm are available for text classification in machine learning. Support Vector Machines is one of them. We have been chosen this for classification in our experiments. It aligns the text into positive and negative based on the word. It's can handle the large feature. When the problem is linearly separable and set of example are sparse then it handles them robustly. Vector representation used to encode the information collected from the text which provides SVM. And it also provides a good result classify the text related problem.

SVM is classify following this formula:

$$f(x) = \sum_{i=0}^{n}(a_i k(x, x_i) + \mathrm{b}) \quad (2)$$

Where,

$$k(x, x_i) = \exp(-\frac{\|x-x'\|^2}{2\sigma}) \quad (3)$$

C. K-neighbors Classifier

K-nearest neighbor classifier used for classification and regression but there is no need to training for applying this algorithm. It's the commonly used learning algorithm for classification. It's the lazy learning algorithm and non-parametric. So, when a new data set is used for classifying at first it calculates the distance of all point in our data set which is based on K initial value. If K value provides 1 then it merges all of the points which are located nearest distance.

If k>1, It makes a list of K of all data points which represent the minimum distance. It's also made a new data point to follow the maximum data point on the list.

Here used a labeled dataset of two peculiarities like need a new data point for identifying the class which is not labeled data. Then calculate the difference between new data points and all other points. Several distance matrics are available to calculate the distance, it's choosing one of them.

Distance function are following:

Eculidean: $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ (4)

Manhattan: $\sum_{i=1}^{n}|x_i - y_i|$ (5)

Minkowski: $(\sum_{i=1}^{n}(|x_i - y_i|)^q)^{1/q}$ (6)

Where,

n is no of dimensions, x is data point from our data set and y is predicted new data point.

D. Gradient Boosting Classifier

Gradient Boosting is another process for solving the regression and classification problems which are built a tree one by one frequently and then prediction the output adjustment the summation of an individual tree. It has three component like one is a loss function which is optimized, secondly a weak learner which predict output and then thirdly additive model which add a weak learner to the loss function for minimizing the loss. Here weak learners used for gradient boosting and greedy manner constructed by choosing the best split point. Trees can add one at a single time but the model is not to change. Gradient boosting made the gradient boosting method using various parameter like 'min_samples_leaf','max_depth','random_state','n_estimators',' subsample','learning_rate'. Every testing gets the different combination of values in gradient boost classifier model and then evaluate every combination of accuracy get the best result in this classifier.

VII. Experiments and Results

A. Data

At first here collect the total 24178 reviews from amazon product review dataset and we transform the string into a meaning full vocabulary which are detect a review is positive or negative. We get total 16954 meaning full word for classifying our data set.

B. Experiment Setting

Here, we divided our data set into two section, training dataset included 70% of total data and testing data included 30% of total data. Date set distribution show in Table 1.

Table 1: Show the Distribution of Dataset

Total Dataset	Training Dataset	Testing Dataset
24178	16924	7254
1.0	70%	30%

The setting of the classifier is following:

- **NB:** The batch size is 100 and default probability is 0.
- **KNN:** N-neighbors value is 3, leaf_size is 30 and using metric is 'minknwski'.
- **GB:** N_estimators is 100, learning rate is 1.0,max_depth is 1 and random_state is 0.

C. Results

See Table 2 to get the accuracy and got the best accuracy is (98.39%) in NB Classifiers comparison all of the others. Also get the highest precision for Negative (N) is (1.0) for KNN and Positive is (0.99) for KNN and NB classifiers. And the best negative recall is (0.11) for NB and Positive recall is (1.0) for SVC, KNN, and GB. Highest Positive F-measure is (1.0) for GB and Negative is (0.19) for KNN.

So, all classifiers accuracy range from 98.26% to 98.39%. Show the result in different classifier Table: 2

Table 2

Classifier	TP	FN	FP	TN	Accuracy %	Precision (N,P)	Recall (N,P)	F1-Measure (N,P)
NB	13	104	52	7085	**0.9839**	N - 0.20 P - 0.99	N - 0.11 P - 0.99	N - 0.14 P – 0.99
SVC	13	104	52	7085	**0.9826**	N – 0.0 P – 0.98	N - 0.0 P - 1.0	N - 0.0 P – 0.99
KNN	12	105	0	7137	**0.9835**	N – 1.00 P – 0.99	N – 0.10 P - 1.0	N – 0.19 P – 0.99
GB	0	117	0	7137	**0.9826**	N – 0.0 P – 0.98	N – 0.0 P – 1.0	N – 0.0 P – 1.0

D. Conclusion and Future Expect

Sentiment analysis is a popular topic for data analytics and reporting and advanced processing. In this research main goal is to represent the different type of classifiers and compare their result when classifying types of reviews in real life. The proposed approach must more accuracy if we change the parameter in the classifier. The Naive Bayes classifier feature selection for implementing Emphasizing words handling and negation handling.

Finally, the author wishes to work a similar type of work in a different language like Bangla especially. Besides, the author wishes more and more experiment execute in Machine Learning algorithm to increase the accuracy. And also wish to include deep learning to analyze the Bengali language sentiment in the future.

REFERENCES

[1] Zaiuddin, N., & Selamat, A. (2014). Sentiment analysis using Support Vector Machine. *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*. doi: 10.1109/i4ct.2014.6914200

[2] Pang, B., & Lee, L. (2004). A sentimental education. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL 04. doi: 10.3115/1218955.1218990

[3] Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages. *ACM Transactions on Information Systems*, *26*(3), 1–34. doi: 10.1145/1361684.1361685

[4] Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). Movie review mining and summarization. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management - CIKM 06*. doi: 10.1145/1183614.1183625

[5] Chowdhury, S. M. M. H., Abujar, S., Saifuzzaman, M., Ghosh, P., & Hossain, S. A. (2018). Sentiment Prediction Based on Lexical Analysis Using Deep Learning. *Advances in Intelligent Systems and Computing Emerging Technologies in Data Mining and Information Security*, 441–449. doi: 10.1007/978-981-13-1501-5_38

[6] O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., & Smeaton, A. F. (2009). Topic-dependent sentiment analysis of financial blogs. *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion - TSA '09*. doi:10.1145/1651461.1651464

[7] Trilla, A., & Alias, F. (2013). Sentence-Based Sentiment Analysis for Expressive Text-to-Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(2), 223–233. doi: 10.1109/tasl.2012.2217129

[8] Ekawati, D., & Khodra, M. L. (2017). Aspect-based sentiment analysis for Indonesian restaurant reviews. *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*. doi:10.1109/icaicta.2017.8090963

[9] Akhtar, Nadeem, et al. "Aspect Based Sentiment Oriented Summarization of Hotel Reviews." *Procedia Computer Science*, vol. 115, 2017, pp. 563–571., doi:10.1016/j.procs.2017.09.115.

[10] Raghib, Omar, et al. "Emotion Analysis and Speech Signal Processing." *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2017, doi:10.1109/icpcsi.2017.8392246.

[11] Varshney, Vanshika, et al. "Recognising Personality Traits Using Social Media." *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2017, doi:10.1109/icpcsi.2017.8392248.

[12] Bansal, P., Somya, Kamaal, N., Govil, S., Ahmad, T. "Extractive review summarization framework for extracted features" International Journal of Innovative Technology and Exploring Engineering Volume 8, Issue 7C2, May 2019, Pages 434-439

[13] Singh, I., & Sahu, A. K. (2019). A Review on Stone Columns used for Ground Improvement of Soft Soil. *Proceedings of the 4th World Congress on Civil, Structural, and Environmental Engineering*. doi: 10.11159/icgre19.132

