Content Categorization Using Logistic Regression from Bangla Social Media Blog Post

Md. Zubaer Alam
Department of CSE
Brac University
Dhaka, Bangladesh
md.zubaer.alam@g.bracu.ac.bd

Sohel Rana
Department of CSE
Brac University
Dhaka, Bangladesh
sohel.rana@bracu.ac.bd

Md Maksudur Rahaman Sohag
Department of CSE
Brac University
Dhaka, Bangladesh
maksudur.rahaman.sohag@g.bracu.ac.bd

Sheikh Shariful Islam Shimul
Department of CSE
Brac University
Dhaka, Bangladesh
sheikh.shariful.islam@g.bracu.ac.bd

Abstract — **In recent years, various Bangla social media blog platforms play a vital role in day-to-day life due to their ease-of-access, portability and affordability. People have learned to rely on these platforms to make their decision on the area of business, research, education, political etc. But peoples are facing difficulty with these platforms because of the contents are available in the discrete way. So, it is very important to make available the contents in such organize way that they are used these platforms more efficiently considering daily needs. In this paper we introduced the content categorization using Logistic Regression method. Logistic Regression algorithm produced highest score among all algorithms. So, we implemented the algorithm that provides the most reliable performance to classify the Bangla social media blog post with quite proficient in Bangla Language.**

Keywords— Bangla Language, Logistic Regression, Classification, Social Media, Natural language processing (NLP).

I. INTRODUCTION

According to Statistic, around 4.48 billion people are actively using social media worldwide as of 2021, the total number of social media users in Bangladesh is 45 million, according to the report released in February 2021, it is equivalent to 27.2 percent of the total population of Bangladesh. A very large number of data has been comprised over the Internet as a result of enormous dealing with social media platforms which conveys a significant contribution in content analysis. To be specific, analyzing the Bangla social media blog post content by users accumulated from social media contents and posts lead to categorize them into several labels. Posts made on Bangla social media blog are on numerous topics. People's opinions on Bangla social media blog are not structured or labeled. It becomes more complicated to categorize when the posts are in Bangla language. However, we implemented the logistic regression algorithm that provides the most reliable performance to classify the Bangla social media blog post. Our Activities, our selection of words, everything produces some type of information that can be rendered and value extracted from it. With the help of NLP, we can comprehend and even predict human behavior using that erudition. To interpret the essence and structure of sentences, NLP utilizes algorithms. Machine learning for NLP and text analytics requires a set of statistical techniques for recognizing parts of speech, entities and other aspects of the text. By using NLP, we can examine people's feelings and can settle them in different labels. Features had extracted from the text data. By implementing prediction based on the dataset, we dissolved the content of the data and classified into different categories. Our initial purpose is categorizing these opinions from Bangla social platforms to enable searching, filtering, and organizing based on post viewpoint. To accomplish this purpose, we took the support of different NLP libraries like NLTK (Natural Language Toolkit), TF-IDF and various ML toolkit like sci-kit learn, Numpy, Matplotlib, and Pandas. We accumulated raw data from Bangla social media blog platforms then process these raw data and modified them into labeled data. To conclude the whole procedure, we performed logistic regression algorithms. Logistic regression is one of the most important analytic tools in the social and natural sciences. In natural language processing, logistic regression is the baseline supervised machine learning algorithm for classification also has a very close relationship with neural networks. The main motive of this paper is to implement a model that can predict categories (Business, Research, Education, Political etc.) of social media blog posts in Bangla.

The paper incorporates the following sections: section II contains the literature review, section III contains our proposed algorithm IV contains our proposed methodology, in section V comprises the result of our experiment, and section VI includes the conclusion and future work.

II. LITERATURE REVIEW

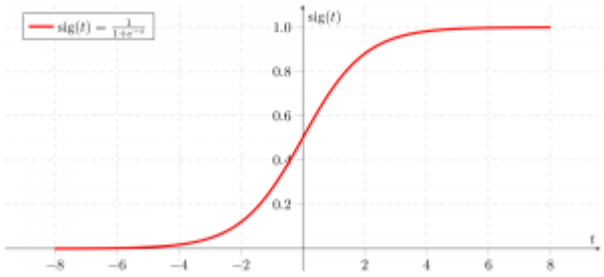
A Bangla news classifier using different ML algorithms is developed by M. S. et. al. They used different classifier algorithms like Naive Bayes, Decision Tree, KNN, SVM, and Random Forest. For measuring accuracy, they used the confusion matrix. After using all of the algorithms they found Naive Bayes has more accuracy than other ML algorithms. The accuracy of Naive Bayes was 85%.
A. K. Mandal et. al. explored the supervised machine learning for the categorization of Bangla web documents. In feature extraction TF-IDF and normalization and finally four classifier algorithms: KNN, SVM, Naive Bayes, Decision tree used. They used 1000 data and for splitting k-fold strategy conducted. One thing they added is training time which calculates the time required for each algorithm and found SVM with the lowest training time.
S. Limon et. al. came with an idea to build up an exceptionally basic site to distinguish or classify the news. In the whole paper, the authors used nine types of data for six news categories which create 6 TSV (tab-separated values) records. These 6 TSV records worked as pre-handled information. They employed ML algorithms like Naive Bayes, SVM, Decision tree, KNN, Random Forest. From all of the algorithms highest accuracy had been given by Naive Bayes which was 76.94% used by them.
M. M. Islam et. al. experimented with the Bangla news headline and approached with a methodology for sentiment analysis on Bangla news headlines. To preprocess data, the authors added extraction and Count Vectorizer for vocabulary count. They classified the news into two sections. With 1600 dataset found the accuracy 75% for SVM, 73% for Naive Bayes.
A way to study online news classification was presented by U. Suleymanov et. al. Azerbaijani news articles were used to collect data. They worked with the dataset and convert textual data into numeric data TF-IDF. They used Kmeans an unsupervised clustering algorithm. SVM and Artificial Neural Network were also used. Artificial Neural Network gave 89% accuracy which was the highest accuracy among all of them.

From the above discussion, we observed that there had been very few works in content analysis in Bangla Language. Comparing the related work, we can perceive that our model provides quite a satisfactory accuracy with a bigger dataset and additionally worked on numerous content categories than others.

III. ALGORITHOM

Logistic Regression – Logistic regression is a supervised classification algorithm. In a classification problem, the target variable y, can take only discrete values for a given set of features, X.
Logistic regression is also called regression model which builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

$$g(z) = \frac{1}{1+e^{-z}}$$



A decision threshold makes logistic regression as classification technique. The setting of the threshold value is a very important matter of Logistic regression and is dependent on the classification problem.

Based on the number of categories, Logistic regression can be classified as:

1. **binomial:** target variable can have only 2 possible types: “0” or “1” which may represent “win” vs “loss”, “pass” vs “fail”, “dead” vs “alive”, etc.
2. **multinomial:** target variable can have 3 or more possible types which are not ordered (i.e., types have no quantitative significance) like “disease A” vs “disease B” vs “disease C”.
3. **ordinal:** it deals with target variables with ordered categories. For example, a test score can be categorized as:“very poor”, “poor”, “good”, “very good”. Here, each category can be given a score like 0, 1, 2, 3.

IV. METHODOLOGY

For applying categorization techniques to Bangla language with the above classifiers, we need to prepare proper datasets for testing and training. Also, like English text classification, pre-processing of Bangla documents and extraction of feature sets are also required before trainings and construction of model for successful document categorization.

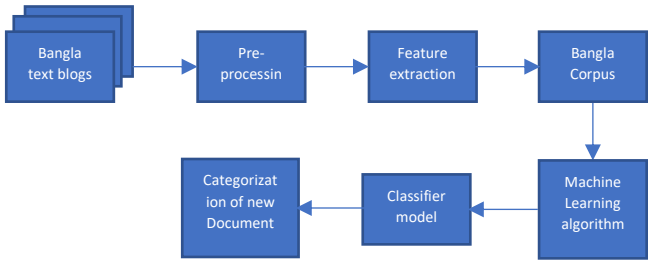


Figure-1: Bangla text classification process

Figure-1 illustrates the overall system of Bangla text classification process which is employed in this project.

A) Proposed Data Sets:
Our paper is about to classify the Bangla words from online blog post where people share their idea in purpose of education, business, or social activities. We will propose in this paper to classify blogs or people comments from Bangla language so that system can detect majority demands, number of positive and negative thoughts.

B) Pre-processing:
It is very important to choose a proper representation of words in text documents for remarkable classification performance. Therefore, the feature extraction or the transforming the input data into the set of features is very urgent. The proper feature extraction will extract the relevant information from the input data to reduce the dimensionality of the feature space and improve the efficiency. But, before extracting features, preprocessing is required. Then following steps are done in pre-processing:

1) Tokenization:
Tokenization is a process which can exchange sensitive data for non-sensitive data called as tokens that can be used in data base without taking advantage. Token has no meaningful value if it violated. Original sensitive information is stored securely outside the organization's internal systems. Tokens serve as reference to the original data but cannot be used to estimate those values. The result of this tokenization step is a set of words limited by white space. Tokenization can be completed in many steps. Suppose a sentence that is “I am a boy”, it can be convert sentences into words tokenization that is “I, am, boy”.

2) Digit Removal:
It is a process which can remove meaningless digits. A simple Bengali text file can contain Bengali as well as English digit. But, a meaningful Bengali word do not accommodates digits, we can remove these meaningless digits using Unicode Representation.

3) Punctuation Removal:
We can remove punctuation and special symbols (^, &, *, (,), | ,:, {,}, [,], etc.). Because, Bengali text file has extra use of spaces, tabs, shifts, etc. For removing punctuation and special symbols, we can use python and java command and online tool. We remove punctuation and special symbols because, it helps to get rid of meaningless parts of the data, or noise. We can remove these by converting all characters to lowercase, removing punctuations marks and typos.

4) Stop words Removal:
It is a process that can remove the words that that are commonly seen in all corpus documents. Stop words do not combine relevant information to the work of text classification. But we have to remove these words from the text document. We compile a list of Bengali language stop words which contains 364 words. There has some one single type letter which has little value. After processing, those single letter are removed in this phase.

5) Stemming:
Streaming is a process which reducing a word to its stem that affixes to suffixes and prefixes. It is important in natural language processing. Suppose, #/##!(ashoner)→#/# (ashon). Word stemming is a major pre-processing process because it acts as a dimension reduction.

6) Feature Extraction:
Feature extraction is the process of converting raw data into the numerical features that can be processed while storing data in the original data

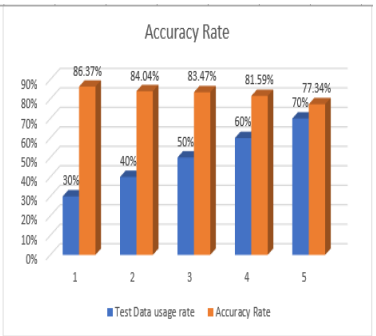
set. After completing the pre-processing phase, we document few numbers of words and extracting of these features from those word. Now it is become easy process. Then we collect the data, these data considered as formal representation. We can call this collection term as corpus. Feature Extraction is possible using different types of statistical approaches. We use TFIDF (Term Frequency–Inverse document Frequency) weighting with length normalization to extracting the features. This method works better than other method for documentation.

V. RESULTS AND DISCUSSION

The dataset comes from the open Bangla blog site repository, and it is related to direct business, research, education, political. We applied Logistic Regression that performed most alike with higher accuracy described that any analysis which is done by using logistic regression as like as using statistical methods. To measure the accuracy of our work, we inserted the accuracy value generated by these algorithms into an accuracy Table. The logistic regression produced the maximum accuracy rate of 86.37% by using 70% training data which is shown in below accuracy Table.

Accuracy Table

Test Data usage rate	*Algorithms*
	Logistic Regression
30%	86.37%
40%	84.04%
50%	83.47%
60%	81.59%
70%	77.34%



VI. CONCLUSION AND FUTURE WORK

In our research, we have developed a model that will categorize the Bangla social media blog posts into different categories based on the textual information. We have implemented Logistic Regression machine learning algorithms and Logistic Regression produced the maximum result. By utilizing the proposed technique our model has been successfully categorized the posts in business, research, education, political. The future guidance on the development of this work is that subcategories of the main category can be established for advanced searching or filtering. To fulfill the purpose, it will necessitate additional datasets. We plan to establish an intelligent system that will automatically categorize the articles or a Blog which can be implemented in different blog sites and social sites dynamically as well as it will classify the textual image.

REFERENCES

[1] D. Chaffey. (17 april 2020). Global social media research summary 2020 Available: https://www.smartinsights.com/social-mediamarketing/
social-media-strategy/new-global-social-media-research/
[2] S. Limon, M. Ahmad, and F. N. Mishu, "Bangla News Classification Using Machine Learning," 2018.
[3] M. S. Salayhin, "Development of a Bangla news classification system,"
2019.
[4] U. Suleymanov, S. Rustamov, M. Zulfugarov, O. Orujov, N. Musayev,
and A. Alizade, "Empirical Study of Online News Classification Using Machine Learning Approaches," in 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, pp. 1-6: IEEE.
[5] L. Nahar, Z. Sultana, N. Jahan, and U. Jannat, "Filtering Bengali Political and Sports News of Social Media from Textual Information," in 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1-6: IEEE.
[6] T. Islam, A. R. Bappy, T. Rahman, and M. S. Uddin, "Filtering political
sentiment in social media from textual information," in 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 663-666: IEEE.
[7] M. M. Islam, A. K. M. Masum, M. G. Rabbani, R. Zannat, and M. Rahman, "Performance Measurement of Multiple Supervised Learning Algorithms for Bengali News Headline Sentiment Classification," in 2019 8th International Conference System Modeling and Advancement
in Research Trends (SMART), 2019, pp. 235-239: IEEE.