Sentiment Analysis on Bangla Language

Akash Bhuiyan[1],Md Ariful Islam[2] and Tasnim Mashrur Mahee[3]
Dept of Computer Science and Engineering
United International University
abhuiyan211068@mscse.uiu.ac.bd[1], mislam211067@mscse.uiu.ac.bd[2],
and tmahee202018@mscse.uiu.ac.bd[3]

***Abstract*—Over the last decade Sentiment Analysis (SA) has become a leading context for scientific and commercial market research, especially in the field of machine learning. It's also called opinion mining. It analyses people's opinion, views, assessments, sentiments towards social units like services, products, organizations, individuals, events etc. Such works are comparatively less witnessed for low-resourced language like Bangla. So, for Bangla language, it is more auspicious research field. In this paper, we propose a framework to analyze and predict the polarity like positive, negative or neutral from a given text. In our work, we have used Bangla comments from social media like facebook and generated a classification model.**

***Index Terms*—Machine Learning, Sentiment Analysis, SVM, Support Vector Machine, Opinion Mining, Bangla Language**

I. Introduction

Sentiment analysis is the process of detecting positive or negative sentiment in text. Since publicly and privately accessible information over the internet is always thriving, people are expressing their thoughts and feelings more openly than ever before, sentiment analysis is becoming an essential tool to monitor and understand that sentiment [1]. It has many practical and empirical applications. Nowadays, the conduct of sentiment analysis is developing day by day in the field of Social Media Monitoring, Brand Monitoring, Voice of Customer, Customer Service, Market Research etc [2].

Sentiment analysis models focus on polarity (positive, negative, neutral) [3]. Depending on how authority wants to interpret people's feedback or opinions. The principal obligation of this sentiment analysis process is to classifying the polarity of a given text. The classification step usually involves a statistical model like Naive Bayes, Logistic Regression, Support Vector Machines, or Neural Networks [4]. Sentiment analysis is one of the hardest tasks in natural language processing (NLP) because even humans struggle to analyze sentiments accurately. The main challenges of SA are Subjectivity and tone, Context and Polarity, Irony and Sarcasm, Comparisons, Emojis, Neutral etc [5].

Bangla is the fourth most popular language in the world and it has approximately 250 million native speakers all over the world [6]. Numerous research works have been done on sentiment analysis for English and other languages such as Chinese, Hindi, Urdu, and Arabic [7], [8]. while sentiment analysis for Bangla is still at a constructive stage. The growing user-generated content on the blogs and news corpus for Bangla forces to monitor the social media and the research for sentiment analysis or opinion mining to get the user opinion of the products or services. We see availability of huge digital contents such as review comments, facebook status written in Bangla which give an opportunity to analyze sentiment in Bangla [9]. However, analyzing this vast amount of data and figuring out individual's interest in a manual way is tedious and in some cases simply inflexible. Thus in these scenarios sentiment analysis can play a vital role to figure out people interests and opinion.

In the recent past, among various methods,recurrent model-based works have enjoyed a lot of success in Natural Language Processing (NLP), compared to more traditional learning methods [10]. While there are other approaches to SA, in this research we will concentrate exclusively on SVM based techniques. Our key contribution cover-

1. A data set of 3000+ Bangla text samples where each sample was annotated by two adult Bangla speakers.
2. Pre-processing the data in a way so that it is readily usable by researchers
3. Application of deep recurrent models on the Bangla text corpus.
4. Pre-train dataset of one label for another (and vice versa) to see if it gives better results

The rest of the part of this paper is organized by the following sections. Section 2 provides a short description of some previous works related to our topic. Section 3 gives an elaborate description of Data collection and Data processing stage. In Section 4, we briefly describe the methodology and discuss the Model Building procedure. In Section 5, the experiment analysis portion has been presented. In Section 6, we conclude the paper and provide some directions for future works.

II. Literature Review

The term sentiment analysis was first introduced by Nasukawa and Yi in their work. They analyse the sentiment by using semantic relationship between sentiment expression and object. The accuracy was 95-90% depending on the data. With the spread of social media on internet, people are getting more opportunity to share their opinion

on digital platform. That is why to deal with this huge data, we observed a similar growth on the research field of sentiment analysis.

However, maximum works related to sentiment analysis are in English language. Due to lack of resources, sentiment analysis in Bangla did not achieved much success. In paper [11] author worked on sentiment analysis on Bangla and Romanised Bangla text. They used the process of Deep Recurrent model, specially Long Short Term Memory (LSTM). They collected data from social media platforms like Facebook, Twitter etc, and some online news portal. The accuracy they obtained is 70%.

In the paper [12] produce a classification model using the neural network variance called Convolutional Neural Network. Their dataset was large enough which contains 120000 data. The positive and negative data ratio was 50%. The accuracy of their model was very high which is 99.87%.

In paper [13] the author used different approach which is contextual valency analysis. Wordnet and SentiWordNet was used to get the parts of speech and prior valance of each word. Then calculated the total positivity, negativity and neutrality of sentence with respect to total sense.

In the paper [14] the author deals with six emotions. The Naive Bayes Classification Algorithm and Topical approach was used to extract the emotions from any Bangla text. The accuracy was almost 80% for Topical approach where the Naive Bayes Algorithm's accuracy was only 60 percent. Both are for 6 classes.

III. Data Collection

The perform ace of any model depends on a suitable dataset. So to produce a great accuracy we need to prepare a good dataset. For better execution we used manually created data corpus. All the data was collected from various Bangla news portal's Facebook posts (Table 1). We collected the Bangla comments from those posts. The dataset consists of total 3043 Bangla text.

TABLE I: Data Sources.

Serial Number	Topic	Sources	Date
1	Crime	BBC Bangla	21.03.2021
2	Others	BBC Bangla	21.03.2021
3	Politics	BBC Bangla	21.03.2021
4	Sports	BBC Bangla	22.03.2021
5	Politics	BBC Bangla	20.03.2021
6	Sports	Prothom Alo	22.03.2021
7	Corona	BBC Bangla	23.03.2021
8	Politics	Ekattor News	25.03.2021
9	Sports	Ekattor News	22.03.2021

To collect data from Facebook we had to design our own crawler. At first our approach was to collect every comment using the Graph Api provided by facebook. But unfortunately this approach did not work as Facebook does not give permission to scrape the public comments. To overcome this situation we used selenium webdriver to collect all the comments from the posts. In this paper we worked on three sentiments- positive, negative and neutral. For this reason we manually set class on each Bangla text. For example, let consider "লোকটি খুব ভালো", this sentence produce a positive sentiment so the class is positive (Table 2). The fig 1 is representing the number of the classes we get in the dataset.

TABLE II: Class for each sentence.

Class	Sentence
pos	সাকিব আল হাসান বাংলাদেশের গর্ব
neg	সাকিব আল হাসান খুব বেয়াদব খেলোয়ার
ntr	সাকিব আল হাসান একজন অলরাউন্ডার

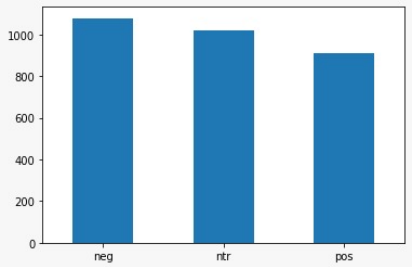

Fig. 1: Class frequency representation

IV. Data Preprocessing

Data preprocessing is transforming the data into a basic form that make it easy to work. Data preprocessing is an important technique for excelling the performance of further used algorithms on that data.

In our paper, we took a total of 3043 Bangla comments as data. The primary collection of comments was noisy and contained URLs, emoticons, punctuations and many unnecessary things. Therefore, we removed the URLs as they do not express any sentiment. Moreover, URLs were removed and sequence of dots between two words and other punctuations were also removed. Many comments contain multiple sentences, thus we placed a comment in one line to indicate them as the parts of a single comment. Furthermore, there were comments written in Banglish (Bangla written in English) and English words. We have removed them to avoid complications.

Fig. 2: Word Cloud

TABLE III: N-gram approach

N-gram	Input	Output
1 gram	মুশফিকের তুলনা হয় না	মুশফিকের, তুলনা, হয়, না
1-2 gram	মুশফিকের তুলনা হয় না	মুশফিকের, মুশফিকের তুলনা,'তুলনা', তুলনা হয়,হয়, হয় না, না
1-3 gram	মুশফিকের তুলনা হয় না	মুশফিকের, মুশফিকের তুলনা, মুশফিকের তুলনা হয়, তুলনা হয় না, 'তুলনা', তুলনা হয়,হয়, হয় না, না

Then we have removed the stopwords. There are many stopwords in Bangla language like অবশ্য, অন্তত, অথবা, অথচ, অর্থাৎ, তাও, সুতরাং, অতএব, এ, ঐ, ইহা, ও, ওই, ইত্যাদি etc.

Then, we have implemented lemmatization so that we can count the appearance of each word. Lemmatization removes the grammar tense and transforms each word into its original form. Such as "পেলো", "পেলাম", "পেয়েছি", "পাইয়াছি", "পেলে", "পেলেও", "পেয়েছিল", "পাইয়াছিল", "পাইলাম", "পেলাম", "পাবো", "পাইবো" etc. words are converted into its original form "পেলো". Like this "করলাম", "করা", "করি", "করেছি", "করেছিলাম" etc. are altered into its original form "কর".

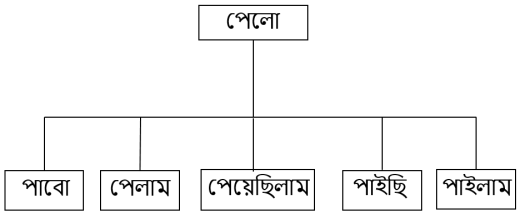

Fig. 3: Lemmatization of word

After applying the above cleansing processes some redundant letters/ alphabets are occurred in the middle of a sentence which does not contain any meaning. We have too removed them too.

The combination of words possess different sense in Bangla, So we applied an approach called N-gram representation which is shown in Table 3. This representation is based on the idea that multiple word terms should contain more semantic information than single words.

Applying Range of N-gram,
for[1] gram, vector dimension = 6790
for[1-2] gram, vector dimension = 25250
for[1-3] gram, vector dimension = 43600

V. Model Building

We have used a total of 3043 data for model building. Among them 80% of data are used as Train Dataset and the rest (20%) are Test Dataset. We have used Support Vector Machine (SVM) technique for model building. SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. These functions can be different types. For example- linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid. The kernel functions return the inner product between two points in a suitable feature space. Thus by defining a notion of similarity, with little computational cost even in very high-dimensional spaces. We have used 3 types kernel for model building – Linear, Polynomial and RBF. Linear kernel is useful when dealing with large sparse data vectors. It is often used in text categorization. Polynomial kernel is usually popular in image processing. RBF is a general-purpose kernel; used when there is no prior knowledge about the data.

We have to set and regulate parameters for each of those kernels.
Linear, SVM regularization parameter = 1
RBF, SVM regularization parameter = 1, gamma = 0.7
Poly, SVM regularization parameter = 1, degree = 3

VI. Experimental Analysis

All the result of the three algorithms are show in the table 4,5 & 6. The linear kernel's result was high among all of them.

TABLE IV: Experimental result of linear kernel

Range of Gram	[1,1]	[1, 2]	[1,3]
Vector Dimension	6790	25250	43600
# of training data	2433	2433	2433
# of test data	603	603	603
Precision	.59	.60	.59
Recall	.60	.59	.58
F-1 Score	.60	.59	.58
Accuracy%	59.8	59.3	58.4

TABLE V: Experimental result of RBF kernel

Range of Gram	[1,1]	[1, 2]	[1,3]
Vector Dimension	6790	25250	43600
# of training data	2433	2433	2433
# of test data	603	603	603
Precision	.61	.69	.70
Recall	.56	.53	.53
F-1 Score	.54	.48	.47
Accuracy%	56.31	53.4	53.15

VII. Conclusion

Based on today's perspective text has become a treasure trove of revealing useful information and people's opinions regarding anything. So uncover the views from the text is an important task now for so many fields like product analysis, social media monitoring, market research and analysis and so on. In this paper, we propose a framework that is

TABLE VI: Experimental result of polynomial kernel

Range of Gram	**[1,1]**	**[1, 2]**	**[1,3]**
Vector Dimension	6790	25250	43600
# of training data	2433	2433	2433
# of test data	603	603	603
Precision	.56	.59	.65
Recall	.50	.45	.40
F-1 Score	.45	.39	.33
Accuracy%	49.8	45.18	40.36

implemented to train a model that can classify sentiment from Bangla comments. Currently, we classify three types of sentiment i.e. Positive, Negative and Neutral. As no Bangla dataset is publicly available, thus for the training purpose we generate a small dataset of Bangla sentences to train and test the model. We collect Bangla comments from different sources, filter the comments and pre-process them to train the network. Due to the lack of smooth Data and Complexity of Bangla Language using supervised method is not an efficient way. Moreover, there still exist some way to improve performance for sentiment analysis from Bangla text. But for that, we need more labeled data which is time consuming. So, in the future, we would try to increase the amount of comments in our dataset to train our model more effectively.

References

[1] Y. Zhang and B. Wallace, ``A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification,'' arXiv preprint arXiv:1510.03820, 2015.

[2] S. D. Tembhurnikar and N. N. Patil, ``Sentiment analysis using lda on product reviews: A survey,'' International Journal of Computer Applications, vol. 975, p. 8887, 2015.

[3] K. Hasan, A. Mondal, A. Saha et al., ``Recognizing bangla grammar using predictive parser,'' arXiv preprint arXiv:1201.2010, 2012.

[4] M. A. Islam, K. A. Hasan, and M. M. Rahman, ``Basic hpsg structure for bangla grammar,'' in 2012 15th International Conference on Computer and Information Technology (ICCIT). IEEE, 2012, pp. 185--189.

[5] B. Liu, ``Sentiment analysis and opinion mining,'' Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1--167, 2012.

[6] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter, and A. K. Das, ``An automated system of sentiment analysis from bangla text using supervised learning techniques,'' in 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS). IEEE, 2019, pp. 360--364.

[7] L.-W. Ku and C.-W. Sun, ``Calculating emotional score of words for user emotion detection in messenger logs,'' in 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI). IEEE, 2012, pp. 138--143.

[8] P. Pandey and S. Govilkar, ``A framework for sentiment analysis in hindi using hswn,'' International Journal of Computer Applications, vol. 119, no. 19, 2015.

[9] M. H. Alam, M.-M. Rahoman, and M. A. K. Azad, ``Sentiment analysis for bangla sentences using convolutional neural network,'' in 2017 20th International Conference of Computer and Information Technology (ICCIT). IEEE, 2017, pp. 1--6.

[10] Y. LeCun, Y. Bengio, and G. Hinton, ``Deep learning,'' nature, vol. 521, no. 7553, pp. 436--444, 2015.

[11] A. Hassan, M. R. Amin, A. K. Al Azad, and N. Mohammed, ``Sentiment analysis on bangla and romanized bangla text using deep recurrent models,'' in 2016 International Workshop on Computational Intelligence (IWCI). IEEE, 2016, pp. 51--56.

[12] M. H. Alam, M.-M. Rahoman, and M. A. K. Azad, ``Sentiment analysis for bangla sentences using convolutional neural network,'' in 2017 20th International Conference of Computer and Information Technology (ICCIT). IEEE, 2017, pp. 1--6.

[13] K. A. Hasan, M. Rahman et al., ``Sentiment detection from bangla text using contextual valency analysis,'' in 2014 17th International Conference on Computer and Information Technology (ICCIT). IEEE, 2014, pp. 292--295.

[14] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter, and A. K. Das, ``An automated system of sentiment analysis from bangla text using supervised learning techniques,'' in 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS). IEEE, 2019, pp. 360--364.