2019 International Conference on Bangla Speech and Language Processing (ICBSLP) 978-1-7281-5241-7/20/$31.00 ©2020 IEEE 10.1109/ICBSLP47725.2019.201497

Data Set For Sentiment Analysis On Bengali News Comments And Its Baseline Evaluation

Md. Akhter-Uz-Zaman Ashik
Department of CSE
Shahjalal University of Science and Technology,
Sylhet,Bangladesh
Email:ashikchowdhury76@gmail.com

Shahriar Shovon
Department of CSE
Shahjalal University of Science and Technology,
Sylhet, Bangladesh
Email: shahriarshovon69@gmail.com

Summit Haque
Department of CSE
Shahjalal University of Science and Technology,
Sylhet, Bangladesh
Email: summit.haque@gmail.com

***Abstract*—The biggest challenge of Bengali language processing is creating a strong data set to do research on. The main focus of this paper is to introduce an authentic and credible data set and this dataset is open for all to be used for educational purposes[1] for Bengali sentiment analysis where the data was extracted from a well known online news portal's user comments. Here comments on various news were scraped, and for detecting the true sentiments of the sentences, five labels of sentiments were used. An online crowd sourcing platform was used for data annotation. To ensure the credibility and validity of the data set, every entry of the data set was tagged three times. Three models of text classification were used for baseline evaluation to check the validity of the data set. This data set might be of valuable help for future works and researches on Bengali sentiment analysis.**

Keywords:: **Sentiment Analysis, Data Set, Bengali, News Comments, SVM, RNN, LSTM, CNN.**

I. Introduction

Peoples value and opinion have a strong impact in today's world. In a world where almost everything is online, we get a large amount of feedback from people in these online sites, blogs and forums. Sentiment Analysis is the method to get a comprehension of this huge amount of opinion from people in a technological way. SA(sentiment analysis) has revolutionized the way we perceive data and make a decision out of this growing amount of data.

People like to express their opinion in online sites, and preferably in their own language. Bengali is a very widely spoken language. Many Bengali news portals have gained popularity in recent decade. SA is fairly very new to Bengali language as opposed to English.

A. Motivation

Language is often vague or highly contextual which makes it very difficult for a machine to understand without human help. As such, human annotated data is essential when training a machine learning platform to analyze sentiment. The performance of a machine learning model to detect sentiments depends largely on the training data. For this reason three different perspectives have been used to make the data set more precise.

There are plenty of scopes to do analysis and research on the aspect of sentiment analysis on Bengali language. Countless news portals, social sites, blogs, etc. have been on the rise which are using this language. These online sites can throw valuable insight on the peoples sentiment as a whole. One of the major challenge of performing sentiment analysis on Bengali language is creating a authenticated and credible data set without ambiguous data. If the data is classified under the opposite category by two different person while preparing the data set, then we can call it 'ambiguous'. The data set on which the sentiment analysis is to be done has to be free of these sort of data. One other thing that inspired us was when a data is labeled according to a persons opinion, that person's opinion is not judged whether s/he is right or not. So we decided to create a data set where the data set will not be labeled according to just one person's opinion to ensure the credibility of the data set.

II. Related Work

Sentiment Analysis means the characterization of the sentiment content of a text unit using Natural Language Processing, statistics or Machine Learning methods. The works of Minqing Hu and Bing Liu circa 2004 done on customer reviews was the first major work done on sentiment analysis. They proposed the Feature-Based Opinion Mining Model which is now known as Aspect-Based Opinion Mining[1]. Md. Atikur Rahman and Emon Kumar Dey presented a work based on their data sets for ABSA(aspect based sentiment analysis)[2]. Their data set had around 5092 data in total. They worked with three tags positive, negative, neutral. M. Nabil et al created a data set from Arabic sentiment tweets consisting of 10,000 tweets and did experiments with 4 way sentiment classification[3]. Akhtar, Md Shad et al created a data set for

[1]Data set: https://data.mendeley.com/datasets/n53xt69gnf/3

978-1-7281-5241-7/19/$31.00 ©2019 IEEE

aspect based sentiment analysis in Hindi and did it's baseline evaluation for data validation where the data set contains 5,417 review sentences across 12 domains. There are a total of 2,290 positive, 712 negative, 2,226 neutral and 189 conflict reviews [4]. A. K. Paul and P. C. Shill did sentiment analysis using mutual information for feature selection and multinominal Naive Bayes for classification using English language data, they have achieved 85.1% accuracy without using negation and got 85.8% accuracy with negation using English testing data. For Bengali, using Bengali testing data, they got 84.78% accuracy without using negation and got 83.77% accuracy with negation[5]. Sentiment analysis was also done in micro blog posts by S. Chowdhury and W. Chowdhury. They used support vector machine and maximum entropy to do the comparative analysis of these two machine learning algorithms[6]. M. S. Islam et al presented a research paper where six different approaches were discussed to evaluate the sentiment. They also discussed about the implementation of cosine similarity using TF-IDF to determine sentiments more accurately[7][8].

III. Methodologies

A. Data Collection

Online news papers have a huge collection of user comments. The data collected was from a widely popular online news portal Prothom-Alo's user comments [2]. From this huge collection of comments 10 specific fields were selected. There are other fields where users comment expressing their opinion but compared to these 10 fields they are negligible. These 10 fields were selected as these are fairly common in comments.

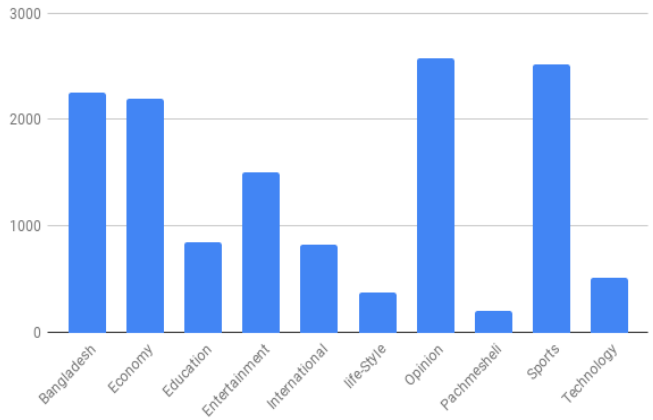

Fig. 1. Categories of the comments

Priorities were given to "Bangladesh", "Economy", "Opinion" and "Sports" because of being the most occurring theme among the comments.

[2]prothomaalo.com

B. Data Pre-processing

The data which was crawled had some characteristic problems within it. The issues we faced were:

- Sentences got divided and separated while crawling
- Some sentences were not properly structured
- Same sentences appeared more than once
- Some sentences had signs and unnecessary characters

To solve these problems the sentence which was not properly structured was dropped. Some signs like multiple question marks, multiple dots and exclamatory signs were cleared. The comments which appeared more than once were dropped.

C. Data Annotation

We wanted to create our own data set with proper attention given to the credibility of the sentiment behind the sentence. When a sentence is labelled, a single tag cannot ensure the actual sentiment of the sentence. A sentence might seem negative to an individual but might not be for another individual. The data set has each entry tagged by three different individuals to get three different perspectives.

There are many approaches to data labelling. Among them, crowd-sourcing is a convenient approach for sentiment analysis. Crowdsourcing is a practice in which information or inputs are obtained from a huge number of people, typically via the Internet. We used this practice for data annotation. Pipilika's crowdsourcing platform [3] was used for campaigns. The data set consists of the standard 5 category sentiments which are strongly positive, positive, neutral, negative, strongly negative where every sentence is labelled 3 times by 3 different individuals to ensure credibility.

D. Processing

Every entry having been tagged three times, they had to be processed and turned into one single tag. So we decided to assign a particular value against every tag. Then by calculating these values we decided to choose the final tag. Suppose an entry has three tags, one positive and two negative, as negative tag appears the most, that entry was labelled to be negative. There were some data where the data was tagged with three different labels which makes the data ambiguous. So we decided to drop entries like this. As a result the data set is free from ambiguous data.

E. Data Set Statistics

The data set we built has 13809 entries in it. It has mainly 5 Labels of Sentiment:

[3]crowd.pipilika.com

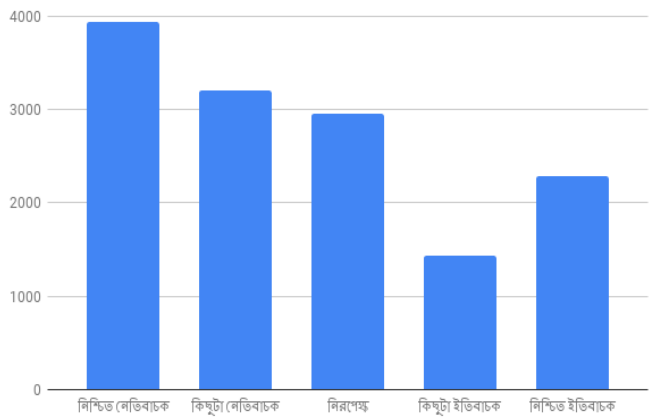

Fig. 2. Labels and there amount

TABLE I
LABEL TYPE AND COUNT

Label	Count
Slightly Positive	1436
Positive	2279
Neutral	2955
Negative	3936
Slightly Negative	3203

An overall map of the percentage of the sentiments is given below:

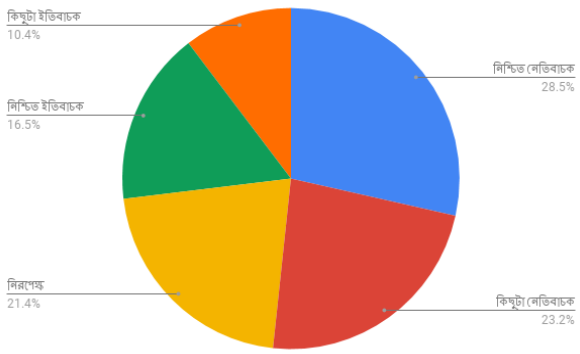

Fig. 3. Percentage of each labels

The data set does not have a negative or positive bias. The conventional model evaluation methods do not accurately measure model performance when faced with imbalanced data-sets. Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes which have number of instances. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of wrong classification of the minority class as compared to the majority class[9].
The statistical summary of the data set:

TABLE II
DATA SET STATISTICS

Category	Words
Total	248562
Longest Sentence	118
Average Sentence Length	44
Numeric Words	2389
Bengali Words	244432
Non-Bengali Words	4130

The various topics we tried to cover are given below in the table:

TABLE III
TOPICS OF THE DATA SET

Fields of Data	Number
Opinion	248562
Sports	118
Bangladesh	44
Economy	2389
Entertainment	244432
International	4130
Education	4130
Technology	4130
Lifestyle	4130
Various	4130

Opinion, Sports, Bangladesh and Economy are the majority of the topics covered.

IV. BASELINE EVALUATION

There are mainly three ways for classifying the sentiment of a text unit[10]. These are Machine Learning Techniques, Lexicon-Based Techniquesand and Hybrid Techniques.

Sentiment Analysis uses the evaluation metrics of Precision, Recall, F-score, and Accuracy for the classification problem. Also, average measures like macro, micro, and weighted F1-scores are useful for multi-class problems. Based on the balance of classes of the data-set the appropriate metric should be chosen. Four effective measures have been selected for the study based on the confusion matrix of the output. These are:

$$Precision(P) = TP/(TP + FP) \quad (1)$$

$$Recall(R) = TP/(TP + FN) \quad (2)$$

$$Accuracy(A) = (TP+TN)/(TP+TN+FP+FN) \quad (3)$$

$$F1 - Score(F_1) = 2.(P.R)/(P + R) \quad (4)$$

Our focus point here is performing Baseline Evaluation on the data-set that we have created on the Bengali news comments for sentiment analysis, where we have selected three models for the Baseline Evaluation. These are: Binary SVM classifier, Multi-class SVM classifier and LSTM(Neural Network).

A. SVM classifier

In machine learning, support-vector machines are supervised learning models that have associated learning algorithms which analyze data used for classification and regression analysis[11]. We can use an SVM classifier when the data has exactly two classes. An SVM classifies data by finding the best hyperplane that separates all the data points of one class from those of the other class.

The data-set was split into test data and train data in the following way:

- Amount of Train set: 10809
- Amount of Test set: 3000
- Amount of Validation Set: 2809

SVM classifier is pretty good for binary classification, when the number of categories is only two, negative and positive. Taking the *slightly positive* and *positive* classes into the *positive* label and the rest into the *negative* label, and dropping the *neutral* class

Following is the confusion matrix we have generated using the SVM classifier:

TABLE IV
CONFUSION MATRIX OF BINARY SVM CLASSIFIER

	Predicted Positive	**Predicted Negative**
Actually Positive	2823	892
Actually Negative	1714	5425

The following table shows the overall accuracy:

TABLE V
ACCURACY OF BINARY SVM CLASSIFIER

Model	**Test(%)**	**Train(%)**	**Validation(%)**
Binary SVM classifier	66.232	93.397	67.488

If we take the *Precison*, *Recall*, *Accuracy* and *F1-Score* into consideration of this model, we generate the table:

TABLE VI
EVALUATION OF THE SVM MODEL

Precision	**Recall**	**Accuracy**	**F1-score**
57.59	72.41	61.34	63.97

B. RNN(LSTM)

Recurrent Neural Networks(RNN) and Long Short-Term Memory(LSTM) networks are often used for sentiment analysis. Recurrent Neural Networks and Long Short-Term networks introduces a memory into the model. Having a memory in a network is useful because, when dealing with text data, the meaning of a word depends on the context of the previous text. A drawback of the Recurrent Neural network is that it is only capable of dealing with short-term dependencies. Long Short-Term Memory networks address this problem by introducing a long-term memory into the network[12]

We created the training , testing and validation splits as follows:

- Amount of Train set: 11041
- Amount of Test set: 1380
- Amount of Validation set: 1381

Like the SVM classifier, we took the *slightly positive* and *positive* classes into the *positive* label and the rest into the *negative* label, and dropping the *neutral* class, we did the evaluation.

The confusion matrix came up to be the following:

TABLE VII
CONFUSION MATRIX OF LSTM CLASSIFIER

	Predicted Positive	**Predicted Negative**
Actually Positive	2529	1151
Actually Negative	1495	5867

The following table shows the accuracy of this model:

TABLE VIII
ACCURACY OF THE LSTM MODEL

Model	**Test(%)**	**Train(%)**	**Validation(%)**
LSTM	74.741	96.967	78.833

By using the measurements of *Precison*, *Recall*, *Accuracy* and *F1-Score*, we can generate the table:

TABLE IX
EVALUATION OF THE LSTM MODEL

Precision	**Recall**	**Accuracy**	**F1-score**
72.84	87.22	74.74	79.291

C. CNN

Convolutional Neural Networks(CNN) is a class of Deep Neural Networks. CNNs are multi-layer perceptrons that are regularized. Multi-layer perceptrons usually have fully connected networks, which can cause over fitting. CNNs solve this issue by taking advantage of hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns.

For text classification with CNN, we usually embed the words of a sentence into a *2D* array stacking them together. Convolution filters are applied to selective number of words to produce a new feature representation. Then some pooling is performed on new features, and the pooled features from different filters are concatenated with each other to form the hidden representation. These representations are then followed by one (or multiple) fully connected layer(s) to make the final prediction[13].

We split the data in the following way for the evaluation:

- Amount of Train Data: 10809
- Amount of Test Data: 3000
- Amount of Validation Data: 2809

The confusion matrix for CNN:

TABLE X
CONFUSION MATRIX OF CNN CLASSIFIER

	Predicted Positive	**Predicted Negative**
Actually Positive	2144	1373
Actually Negative	1862	5455

The accuracy table for CNN:

TABLE XI
ACCURACY OF THE CNN MODEL

Model	**Test(%)**	**Train(%)**	**Validation(%)**
CNN	60.49	89.03	63.68

The evaluation of the CNN model:

TABLE XII
EVALUATION OF THE CNN MODEL

Precision	**Recall**	**Accuracy**	**F1-score**
58.92	68.52	60.49	66.24

V. DISCUSSION

SVM employs kernel tricks and maximal margin concepts to perform better in non-linear and high-dimensional tasks. SVMs are great for relatively small data sets with fewer outliers.

Neural networks typically perform better on very large data-sets. Neural networks profit a lot if the data points are structured in a way that can be exploited by the architecture. That is the case with our data-set, thats why neural networks (LSTM) is a better choice. Neural Networks may require more data but they almost always come up with a pretty robust model.

Deep learning really shines when it comes to complex problems such as image classification, natural language processing, and speech recognition. The data set has no clear distinct pattern which can be exploited by the CNN model.

VI. CONCLUSION

In our endeavour, we have created our very own data set for analysis, which consists of 13809 entries from the Prothom-Alo news portal. Baseline model selection and evaluation have been done on this data set using three different models which are SVM, CNN and LSTM model. We have done some comparative performance testing with our data set related to Bengali sentiment analysis and presented them in a tabular form.

ACKNOWLEDGMENT

We are thankful to our Department of Computer Science and Engineering and NLP group of Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh, for the help they have always provided us with and for motivating this research.

REFERENCES

[1] B. Liu, "Opinion mining, sentiment analysis, opinion extraction," available at: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. (Accessed on 16 May 2019).

[2] M. A. Rahman and E. Kumar Dey, "Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation," *Data*, vol. 3, no. 2, 2018. [Online]. Available: https://www.mdpi.com/2306-5729/3/2/15

[3] M. Nabil, M. Aly, and A. Atiya, "Astd: Arabic sentiment tweets dataset," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2515–2519.

[4] M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Aspect based sentiment analysis in Hindi: Resource creation and evaluation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2703–2709. [Online]. Available: https://www.aclweb.org/anthology/L16-1429

[5] A. K. Paul and P. C. Shill, "Sentiment mining from bangla data using mutual information," in *2016 2nd International Conference on Electrical, Computer Telecommunication Engineering (ICECTE)*, Dec 2016, pp. 1–4.

[6] S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in bangla microblog posts," in *2014 International Conference on Informatics, Electronics Vision (ICIEV)*, May 2014, pp. 1–6.

[7] M. Al-Amin, M. S. Islam, and S. Das Uzzal, "A comprehensive study on sentiment of bengali text," in *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Feb 2017, pp. 267–272.

[8] M. S. Islam, M. A. Amin, and S. Das Uzzal, "Word embedding with hellinger pca to detect the sentiment of bengali text," in *2016 19th International Conference on Computer and Information Technology (ICCIT)*, Dec 2016, pp. 363–366.

[9] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," *arXiv preprint arXiv:1305.1707*, 2013.

[10] S. Symeonidis, "5 things you need to know about sentiment analysis and classification," available at: https://www.kdnuggets.com/2018/03/5-things-sentiment-analysis-classification.html. (Accessed on 16 June 2019).

[11] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[12] F. Miedema, "Sentiment analysis with long short-term memory networks," *VRIJE UNIVERSITEIT AMSTERDAM*, vol. 1, 2018.

[13] S. Minaee, E. Azimi, and A. Abdolrashidi, "Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models," *arXiv preprint arXiv:1904.04206*, 2019.