2019 22nd International Conference on Computer and Information Technology (ICCIT),
18-20 December 2019

Employing Machine Learning techniques on Sentiment Analysis of Google Play Store Bangla Reviews

Md Muhtasim Jawad Soumik
dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
muhtasimjawad007@gmail.com

Syed Salvi Md Farhavi
dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
salvifaravi111@gmail.com

Farzana Eva
dept. Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
farzana0023@gmail.com

Tonmoy Sinha
dept. Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
tonmoy2101@gmail.com

Mohammad Shafiul Alam
dept. Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
shafiul.cse@aust.edu

***Abstract*—This article offers an in-depth insight on a number of existing methodologies to perform sentiment analysis using text classification on Bangla dataset. Although the rapidly developing machine learning algorithms are showing promising results, the viability of those methods for non-English languages such as Bangla is yet to be fully explored. This research aims to fill in some of those existing research gaps through proper implementation of machine learning techniques where words are converted into feature vectors via implementation of TF-IDF algorithm on data crawled from Google play store, the largest Android application market. Many significant algorithms staring linear algorithms like Naïve Bayes, Linear Support Vector Machine (SVM) are implemented. An in-depth comparison is also made among the results of various existing algorithms. The experimental results indicate that even the base-line algorithms, after proper pre-processing, can show promising results on our Bangla dataset. Naïve Bayes, Support Vector Machine and Logistic Regression has shown very promising results (accuracy score of 0.75 on average) even with the data limitation. An Ensemble method is also proposed with Adaptive Boosting technique showing an accuracy score of 0.7639 with five-fold applied. SVM has the best accuracy score of 0.7648 among all the algorithms when five-fold is applied and Gradient Boosting has the best accuracy score of 0.7695 when five-fold is not applied.**

***Index Terms*—sentiment analysis, natural language processing, machine learning, google play store, ensemble method**

I. Introduction

To be one step ahead of others in this is highly competitive world of marketing, it is an absolute necessity to understand user sentiment towards a specific product and remodel/refine it accordingly. As the main goal of any production company is to sell more and more products to the target buyers/audience, what users want holds the most importance to them. Today users are constantly looking for products to match their ever-changing taste which is more true for an open market like the market for android apps also known as Google Play Store. As it stands, most of the applications are free to download and users can get vast number of choices for a specific type of application which means more and more competition among the developers which again brings us to importance of the quality control for the applications, understanding user sentiment and updating applications accordingly. Every application page on the Google play store has a comment section where the users convey their constructive criticism. Now it is rather obvious that only a human may come close to realizing the sentiment of another human through the written text but it is impossible to do so manually as an app can have millions of users and also millions of comments. So that machine learning can play a significant part here as there is no shortage of data and it can be a great way to improve the applications. Sentiment Analysis using Machine learning has already yielded very

978-1-7281-5842-6/19/$31.00 ©2019 IEEE

promising results in user review based scenarios such as: IMDb (Internet Movie Database), Amazon store review and so on but despite of being a very popular application market place, the lack of research on this field can be seen. One major challenge to conduct sentiment analysis on the play store reviews is the language barrier. With the advancement of research based on machine learning and sentiment analysis, different libraries can be found which can simplify the whole process of text pre-processing to classification of text for English language. To close down this language barrier, text on Bangla language is chosen not only for it being the national language of but also for the fact that accounding to stat Counter[1](Feb,2019) android holds 75 percent of total operating system market share in Bangladesh. Moreover, Bangla is a widely spoken language currently sitting at the fourth position based on the number of people speaking Bangla as their first language[2].

This paper shows the process followed to perform sentiment analysis successfully starting from data fetching from playstore to applying various machine learning algorithms and result comparison among the algorithm implementations.

II. Related Work

Sentiment Analysis is a popular topic to work on and many projects or researches done on movie review, twitter data, product review etc.

Microblog posts like tweets are used for classifying sentiment by Phani, Lahiri and Biswas (2016). They tried to work with three different languages (Bangla, Tamil and Hindi). They performed stratified 10-fold cross validation on the training data [5]. For cross validation, they experimented with word n-grams, character n-grams, surface features and Sentiword features. Agarwal, Xie, Vovsha, Rambow, Passonneau (2011) also used twitter data and they introduced POS-Specific features which is based on polarity. To perform the 3-way classification of tweets (positive, negative and neutral), they chose to work with three types of model: Unigram model, a feature based model and tree kernel based model [3].

Sentiment analysis also done in epinions review by Turney (2002). Author used unsupervised learning algorithm to identify sentiment and predicted a review by average semantic orientation of the phrases in the review that contains adjective or adverb. The Pointwise Mutual Information and Information Retrieval (PMI-IR) algorithm is used to estimate semantic orientation of the phrases [2]. Kiritchenko, Zhu, Cherry and Mohammad (2014) determined aspect terms, aspect categories and sentiment from customer reviews. They used PMI method to create sentiment lexicons and Brown clustering algorithm to create word clusters. They used semi-Markov tagger to tag token sequence and trained the tagger using the structural Passive-Aggressive (PA) algorithm. They divided their features into two categories: emission and transition features. They used multi-class SVM algorithm to classify sentiment (positive, negative, neutral and conflict) [4]. Nguyen, Nguyen and Pham (2013) used Naïve Bayes and SVM to classify the sentiment in two stages. At first, they used naïve bayes classifier to determine the sentiment. They forwarded the misclassified sentiments from naïve bayes to the second stage which was SVM to classify those misclassified sentiments [1]. Tanmoy Chakraborty and Sivaji Bandyopadhyay (2010) identified the reduplications at expression and semantic level in Bengali. They identified the MWEs (multiword expression) at tokenization phase and then POS tagger identified those words as unknown words. Bengali Shallow Parser was used to identify the hyphened reduplications. They designed the system in two phases where the first phase identified five cases of reduplications (complete, partial, onomatopoeic, correlative and semantic) and the second phase attempted to extract the associate sense [6].

III. Data Description

Google Play store represents different types of android applications, most of which come free. So the whole review section is a reflection of opinions from people of different tastes and mentality. A web crawler was built from the scratch for the data collection purpose which yielded critical review information: user name, rating, review body. Some of the challenges were visible from the early stage of data collection such as misspelling, lexical variation, slangs, emoticons. Figure 1 shows working processes followed to collect Bangla dataset.

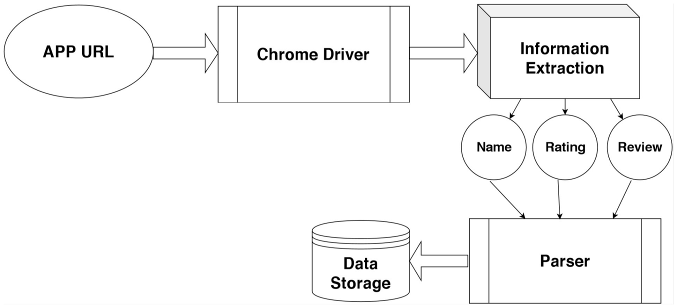

Fig. 1. Process of dataset creation

Selenium and Beautiful Soup are the library and the parser which were used for automated crawling of reviews. Reviews are collected from more than 100 apps (Bkash, Daraz, Uber etc.) and acquired around 10000 reviews and 6500 unique words.

[1] http://gs.statcounter.com/os-market-share/all/bangladesh
[2] https://www.ethnologue.com/13/top100.html

TABLE I
Preview of Dataset

Name	Rating	Review
Sharmin Akter	5	চমৎকার অ্যাপ। আরও নতুন বই চাই।
রমজান হোসেন	1	কেউ ইনস্টল করবেন না। এটাতে শুধু এড আর এড। খুবই ফাউল অ্যাপ। এত বাজে অ্যাপ আমি আগে দেখি নাই।
Md. Sakib	3	সবকিছু ভাল লেগেছে। কিন্তু পেজ নাম্বার দেয়া নাই। ফলে কত পেজ পড়া হয়েছে তা বুঝা কষ্টকর।

Reviews are annotated manually by human annotators as positive, negative and neutral and denoted as 3, 1 and 2 respectively. Few examples showing result of the annotation process are shown below.

TABLE II
Preview of Dataset after Annotation

Review	Annotation
চমৎকার অ্যাপ। আরও নতুন বই চাই।	3
কেউ ইনস্টল করবেন না। এটাতে শুধু এড আর এড। খুবই ফাউল অ্যাপ। এত বাজে অ্যাপ আমি আগে দেখি নাই।	1
সবকিছু ভাল লেগেছে। কিন্তু পেজ নাম্বার দেয়া নাই। ফলে কত পেজ পড়া হয়েছে তা বুঝা কষ্টকর।	2

Equal amount of reviews are taken for generating unbiased result and the dataset is divided into five parts for cross-validation purpose.

IV. Methodology

The steps involved in Sentiment Analysis are:

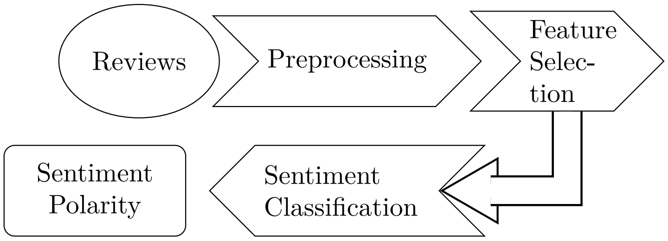

Fig. 2. Process of Sentiment Analysis

A. Preprocessing

All the reviews are pre-processed as follows:

- With the help of WhitespaceTokenizer() method, tokens were extracted from string of words or sentences without whitespaces, new line and tabs
- Words in each sentence were sorted according to the standard sorting order defined by Bangla Academy[1]
- A list containing 398 Bangla stopwords was taken from a github repository and more words were added to the list and used for removing stopwords[2] was used for common stop word detection and removal
- Each and every word in a review is not that significant, on the other hand certain words show their own weightage by occurring number of times. This was achieved through frequency distribution

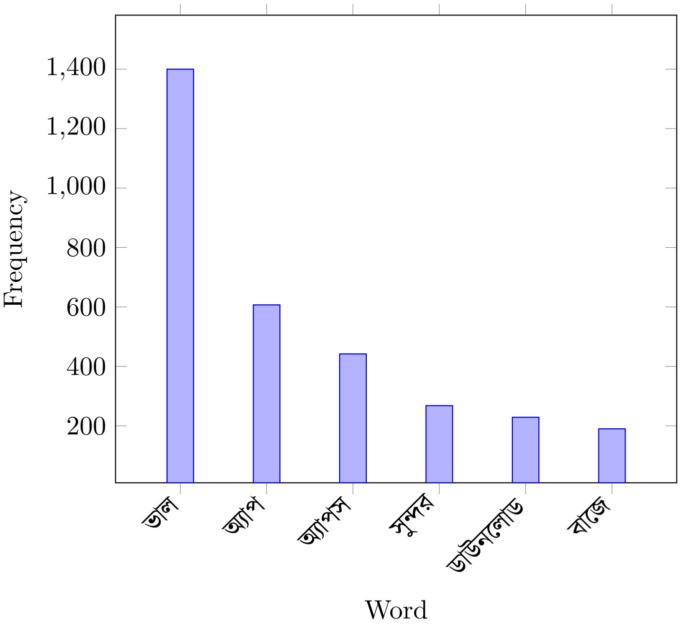

Fig. 3. Frequency distribution of words

- Manual correction was done for spelling mistakes and lexical variations due to lack of any other efficient process for Bangla

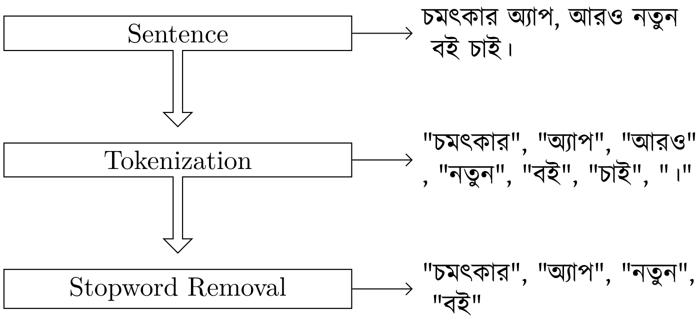
Sentence
চমৎকার অ্যাপ, আরও নতুন বই চাই।
Tokenization
"চমৎকার", "অ্যাপ", "আরও" , "নতুন", "বই", "চাই", "।"
Stopword Removal
"চমৎকার", "অ্যাপ", "নতুন", "বই"

Fig. 4. Preprocessing work flow

B. Feature Selection

Feature in language processing scope refers to the numeric vectors converted from textual data. Since the textual data here is of different language, building up a feature selection system from scratch seemed to the most convenient solution. To do so, Term Frequency-Inverse Document Frequency (TF-IDF) method is chosen for it

[1] Bangla Academy sorting - https://github.com/banglakit/bangla-academy-sort

[2] Stop word - https://github.com/stopwords-iso/stopwords-bn

being the most popular approach. TF-IDF has tackled the issue of the most frequent words of being undesirable with respect to algorithm implementation by assigning less weight such as the word "অ্যাপ" which is found most frequently in play store reviews but unnecessary for sentiment analysis purpose and is weighted close to zero.

C. Classification Techniques

The dataset is a completely new one so, no external dataset was available for testing. For this reason the dataset was divided into train set and test set and all the algorithms were applied on them. Five fold cross validation was applied to get unbiased result as there is no guarantee that changing the test set will not yield accuracy score less than before. The techniques which are implemented are as follows:

a) Naïve Bayesian Classifier: In natural language processing (NLP) problems, naïve bayes classifier is widely used. Multinomial naïve bayes is used in this paper. This classifier calculates the probability of each tags of a document and results the highest one. It works well for data which can be easily turned into counts, such as word counts in text and TF-IDF vectorizer is being used to turn words into numbers which involves word count.

b) Support Vector Machine: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. The vectors (cases) that define the hyperplane are the support vectors. Linear Support Vector Machine which has been used in this paper is widely regarded as one of the best text classification algorithms.

c) Logistic Regression: Logistic Regression is a simple and easy to understand classification algorithm and it can be easily generalized to multiple classes. It assumes a linear, additive relationship between the predictions and log odds of a classification. It analyzes a set a data points with one or more independent variables and finds the best fitting model to describe data points using the logistic regression equation. Logistic Regression is very effective for problems in which the set of input variables is well known and closely correlated with the outcome.

d) Ensemble Methods: Ensemble Methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. Which algorithm works best for a certain scenario can not be predicted in most cases before the experimentation. So instead of using only one algorithm and hoping for the best, introducing Ensemble methods can ensure the better algorithm is taken into account by max voting system, averaging the results or by diving into advanced Ensemble techniques like bagging and boosting. The idea behind bagging is to combine the results of multiple models to get a generalized result but if all the models are created on the same set of data, chances are these models will give the same result since they are getting same input. The solution to this problem is bagging/bootstrapping. It is a sampling technique in which subsets of observations from the original dataset are created, with replacement.

A problem can arise where a data point is incorrectly predicted by the first model, then the next which makes combining the predictions to provide better result useless. Such situation can be handled by boosting. This is also a sequential process where each subsequent model attempts to correct the errors in the previous model. Adaptive boosting, Gradient Boosting and extreme Gradient Boosting are tested as all of them are viable option for classification problems with XGBoost already proven to be a highly effective ML algorithm.

V. Experimental Result

After performing Naïve Bayes, Support Vector Machine and Logistic Regression on our dataset which consisted around about 10000 review (positive, negative and neutral) containing about 6500 unique words. The dataset was divided into five parts for cross validation and the bagging and boosting techniques were applied taking Logistic Regression as the baseline algorithm which generated the following results:

TABLE III
Result Comparison

Name of Algorithm	**Accuracy Score**	
	with 5-fold	**without 5-fold**
NB	73.67%	75.98%
SVM	76.48%	75.02%
Logistic Regression	75.87%	75.98%
Bagging Meta-estimator	75.44%	75.81%
Adaptive Boosting	76.39%	75.90%
Gradient Boosting	76.04%	76.95%
Extreme Gradient Boosting	73.03%	72.64%

Here it can be seen that although the naïve bayes does not provide the best result, it is certainly close to other algorithms. This indicates that naïve bayes model works well even in heavy context situations. Next comes the SVM algorithm which shows more promise than naïve bayes. SVM features a few kernel functions among which the simple linear kernel works well and fast. Text is often linearly separable and has a lot of features which justifies using linear kernel for the test. After that, logistic regression was implemented which also showed promise. Implementation of logistic regression has been done solely due to the fact that it is a simple algorithm which can easily be generalized to multiple classes. Lastly, few boosting ensemble methods were tested to see if they can improve the already shown results as the nature of these algorithms is to try to correct the errors of the previous model. The results are not necessarily better than the previous algorithms but good none-the-less.

Among the applied ML algorithms, SVM has the highest accuracy score and has a good performance among them for five-fold. On the other hand, Gradient Boosting gives the best result when done without five-fold.

VI. Limitations and Future Plan

During this long process from data collection to sentiment analysis, not everything went as planned and few obstacles were on the way. Despite of utmost sincerity, this paper also bears some limitation which are as follows:

- Limitation of how much data can be gathered will always be an issue in case of machine learning where more data almost always lead to more accurate learning. So bigger dataset was preferable
- Due to lack of proper toolset for Bangla data preprocessing, the data needed to be cleaned manually which was a hectic work
- During the cleaning process, some of Bangla words were converted into ascii codes which then had to be removed again to prevent further noise in the data
- As dictionary of positive and negative words in Bangla was not found, lexicon-based approach could not be performed

The work done so far is just the beginning of building up a sophisticated sentiment analysis tool for Bangla language focusing on the play store reviews. There were also some ideas which could not be explored due to lack of time and the research group being small. This ideas which are yet to be explored are as follows:

- Gathering more and more data to build up a trustworthy and experimentally proven data which to be made free for public so the research programs regarding this field can take a step forward
- Artificial Neural Network(ANN) is a very popular machine learning technique which for its unique insights and complex calculations also demand a large and well developed dataset. After dataset building is complete, the goal is to use ANN to perform sentiment analysis
- Building up hybrid algorithms combining both basic and complex machine learning techniques which may improve the performance futher

VII. Conclusion

Here a system of playstore sentiment analysis is presented where mainly two approaches: model based on general machine learning algorithms like SVM and ensemble techniques combining different algorithms are followed. Almost all the algorithms showed promising result after the training of dataset which was also created as a part of this research. Before training, the fetched data needed cleaning and proper annotation. Using TF-IDF enabled the option to eliminate feature neutral words. The success of naïve bayes algorithm proves that the sentiment detection system can also ensure good results even without the context although other algorithms have shown better results. So it can be concluded that sentiment analysis on google play store Bangla data resembles other sentiment analysis instances where more data can ensure the use of this research in a broader scope.

References

[1] Dai Quoc Nguyen, Dat Quoc Nguyen and Son Bao Pham 2013. A Two-Stage Classifier for Sentiment Analysis. International Joint Conference on Natural Language Processing, pages 897–901, Nagoya, Japan, October 14-18, 2013.

[2] Peter D.Turney 2002. Thumps Up or Thumps Down? Semantic Orientation Applied to Unsupervised Classification of reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.

[3] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau. 2011. Sentiment Analysis of Twitter Data. In Proceeding of the Workshop on Languages in Social Media, LSM'11, pages 30-38, Stroudsburg, PA, USA, Association for Computational Linguistics.

[4] Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif M. Mohammad 2014. Detecting Aspects and Sentiment in Customer Review. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 437-442, Dublin, Ireland, August 23-24, 2014.

[5] Shanta Phani, Shibamouli Lahiri, Arindam Biswas. Sentiment Analysis of Tweets in Three Indian Languages. Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing, pages 93–102, Osaka, Japan, December 11-17 2016.

[6] Tanmoy Chakraborty and Sivaji Bandyopadhyay. Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule-Based Approach. Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010), pages 73–76, Beijing, August 2010

[7] Dipankar Das, Sivaji Bandyopadhyay. Labeling Emotion in Bengali Blog Corpus – A Fine Grained Tagging at Sentence Level. Proceedings of the Eighth Workshop on Asian Language Resources, pages 47-55, Beijing, China, August 2010.