Comparative Study on Machine Learning Algorithms for Sentiment Classification

Mohammad Mohaiminul Islam
Research Associate, Department of Computer Science & Engineering
Daffodil International University, Bangladesh

Naznin Sultana
Assistant Professor, Department of Computer Science & Engineering
Daffodil International University, Bangladesh

ABSTRACT
Sentiment Analysis is the study of people's opinions and emotional feedbacks towards an entity which can be products, services, individuals or events. The opinions are most presumably be expressed as reviews or comments. With the advent of social networks, forums and blogs, these reviews emerged as an important factor for the customers' decision for the purchase or choice of any item. Nowadays, a vast scalable computing environment provides us with very sophisticated way of carrying out various data-intensive natural language processing (NLP) and machine-learning tasks to analyze these reviews. One such task is text classification, a very effective way of predicting customers' sentiment. This paper investigates the different ways of sentiment analysis from customers' review using machine learning algorithms. For classifying text from overall sentiment, we considered two class, i.e. predicting whether a comment or review is positive or negative. In our study, we used two popular public datasets and six different machine learning algorithms – Naïve Bayes (Multinomial and Bernoulli), Logistic Regression, SGD (Stochastic Gradient Descent), Linear SVM (Support Vector Machine) and RF (Random Forest). Moreover, we applied parameter optimization on SVM and SGD classifiers on different threshold values to identify and analyze the differences in the accuracy of the classifiers and to obtain the optimal outcome from the model.

Keywords
Natural Language Processing, Sentiment Analysis, Opinion mining, Machine Learning.

1. INTRODUCTION
Nowadays sentiment analysis has become one of the popular and interesting tasks for the researchers working in the field of natural language processing. It has become more popular in opinion mining of user's towards products, political reviews, and movie reviews etc. and at the same time we can analyze human sentiments from their posts or comments on web and various social networks sites. Producers, manufacturers, film makers, politicians, health care personnel's can be able to know the views and thoughts of the customers, consumers, viewers and be able to get an idea of a person's mental health by analyzing their reviews and comments from many online sites like facebook, twitter, Orkut, imdb, Amazon etc. The task of sentiment analysis can also be performed in financial services, political influences and other possible domains where humans leaves their opinions on social platforms. Therefore, developed concepts and techniques of information technology can suggest modern solutions that explores text classification with machine learning and works with collections of humans' opinions or customer feedback data expressed by short text messages.

Sentiment Analysis (SA) concerned with the classification of human sentiment into some predefined classes. For this classification task sentiment can be viewed from three abstract levels such as document-level, sentence-level and aspect-level. In this paper we focused on machine learning based sentence level classification task and considered the polarization of sentences into two classes, (i) positive and (ii) negative. We used two different datasets, one is movie reviews collected by crawling from IMDB movie review site [3] and another is Amazon Book review dataset collected from Amazon web site [4]. For classification purpose we choose some popular and widely used supervised machine learning algorithms. The algorithms are Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Stochastic Gradient Descent, Linear Support Vector Machine and Random Forest. We analyzed the performance of these algorithms in different perspectives and finally we came up with a conclusion about the prediction capability of the selected algorithms with the help of some evaluation matrices. The results of our investigation can be used in a variety of large scale textual data processing systems for selecting the model structure and the optimal algorithm based on the nature of the dataset. In addition, our findings will also help data analysts to predict the data to support knowledge gathering and decision support system. The rest of the paper is organized in the following manner: Section 2 provides related works from literature; Section 3 describes the datasets and experimental setup of our model; analysis and comparison of different machine learning algorithms are discussed in Section 4 and Section 5 concludes the manuscript with future extension of this work.

2. RELATED WORKS
This section provides a literature review on sentiment analysis and highlights the major concern of the researchers on their work. Since sentiment analysis is an interesting topic for many researchers so a good number of articles are published every year in this field and the number of articles are increasing through years. Below are some literature review related to our work: according to Pang et al, traditional approaches on sentiment analysis use word count or frequencies in the text which are assigned sentiment value by expert [5]. These approaches disregard the order of the words rather it focuses on the frequency of word. They suggested that a recurrent neural network (RNN) can be used for sequence labeling on sequential data of variable length. According to their study each input sentence fed to the model for sentiment classification is considered to be a collection of tokens. Pang & Lee [5] and Liu [6] conducted a detail survey and the main focus of their survey was on the applications and challenges in SA. Cambria and Schuller et al. [7], Feldman [8], Montoyo and Martı´nez-Barco [9] has provided short surveys in their paper illustrating the new trends in SA.

Tsytsarau and Palpanas [10] also conducted a detail survey covering the major topics of SA. For each topic they have illustrated the definition, problem statement, development process and categorized SA with the aid of tables and graphs. Another more related area of research is that of determining the genre of texts; subjective genres, such as "editorial" is often one of the possible categories [11]. Other works explicitly attempt to find features indicating that subjective language is being used [12].

Some of the work focused on classifying the semantic orientation of individual words or phrases, using a pre-selected set of seed words or linguistic heuristics [13][14]. Some researchers worked on sentiment-based categorization of entire documents which often involved either the use of models inspired by cognitive linguistics [15] or the manual or semi-manual construction of discriminant-word lexicons [16][17]. Turney worked on classification of reviews using unsupervised learning [18]. He focused on the mutual information between document phrases and the two common adjectives "poor" and "excellent". A search engine was used to gather statistics for computing mutual information between document phrases and these two words. However, T Joachim in his work on text classification with supervised machine learning suggested that Support Vector Machine is one of the best classifier compared to that of Naïve Bayes or Decision Tree [19]. Other authors also agreed about the superiority of Support Vector Machine over Decision Tree, and Naïve Bayes [20]. After a thorough review of literature, we decided to make a comparative study on some widely used as well as less investigated supervised machine learning algorithms for text classification to justify the performance of these algorithms in terms of prediction accuracy and other evaluation metrics.

3. DATASET PREPARATION AND EXPERIMENT SETUP

In our study we used two widely used public dataset; IMDB movie review dataset consists of 50K full length reviews on 1500+ movies and Amazon Book review dataset consists of 60K reviews on 9173 individual books. In IMDB dataset there were 25K movie reviews for training and 25K reviews for testing our model. Among them 12.5K reviews were positive and 12.5K reviews were negative. Similarly for test set where 12.5K were positive and 12.5K were negative reviews. In Amazon dataset there were 48K reviews for train set where 24K reviews were positive and 24k reviews were negative. There were 12K reviews for test set where 5911 were positive and 6089 were negative reviews. We randomly selected almost similar numbers of positive-sentiment and negative sentiment to balance out both of our datasets. And for our model, we focused on BOW (Bag of Words) as features selection approach based on unigram. We used Python language to conduct our experiment using Python machine learning library for data and natural language processing. For evaluation purpose we used some matrices like classification Accuracy, Logarithmic loss, Area under ROC curve, Confusion Matrix & Matthews Correlation Coefficient. We established a workflow model for sentiment analysis of text review processing to compare Naïve Bayes (Multinomial & Bernoulli), Logistic Regression, Linear SVM, SGD and Random Forest classifiers. Fig. 1 presents the workflow model for sentiment analysis which is a modified version of that presented by Seddon [21]. The workflow consists of four key stages: Data extraction, Preparation of review texts, Bag of words model and Classification.

3.1 Data Extraction

It is one of the most important preprocessing steps that deals with selecting only the required and related data fields to process in order to optimize memory usage. In our experiment this stage was carried out as follows.

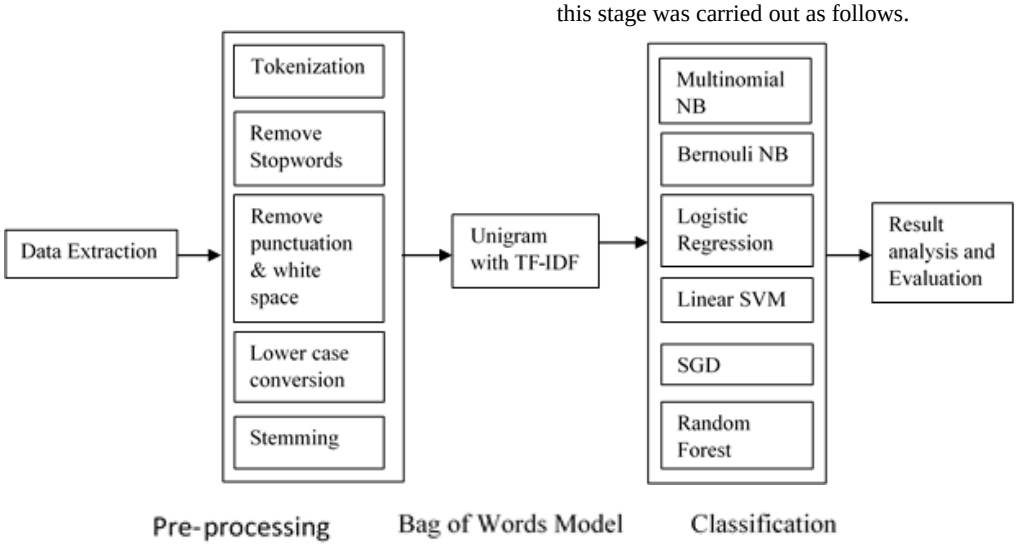


Fig 1: Workflow model for review processing

- Only ratings and review text fields were taken from the input dataset. (IMDB dataset).

- Only ratings, review text, helpfulness and summery fields were taken from the input dataset (Amazon dataset).

- Collecting the equal number of customer product-review records in each class to avoid skewness.

3.2 Preparation of Review Text

This stage is concerned with the preparation of review text and summary fields from the dataset to extract features. Following operations were performed as the data preparation tasks:

- Tokenizing each word of the text and giving an integer id for each possible token by using punctuation or white space as token separators.

- Removing all stop words such as a and the (Stop word corpus was taken from the NLTK website. Stop words a and the are frequently used in any text, but they do not actually carry any specific information required to train the model.

- Converting all the capital letters to a lower case.

- Stemming (with Porter stemmer) and reducing inflectional forms to a stemma form.

- Lemmatizing to group together the different inflected forms of a word so they can be analyzed as a single item.

3.3 Bag of Words Model

It is a process to split the sentence into words and group them using a combination of n-grams. Bags of words (unigrams) are created from review texts that have passed previous stages, based on the unigram model. These words are imported to specially created tfidf that counts the frequency in the set and assigns a unique numerical value for the next classification stage, as well as the weights needed for each word. The feature vector transforms words in to numerical values represented in the integer format, i.e. the numerical value to the given word and the value of frequency of the word.

3.4 Classification

This stage was carried out as follows:

- Data training and testing were performed by the selected classification method using 5-fold cross-validation.

- Calculating the average classification accuracy. Since we aimed to make our experiments repeatable and verifiable, we utilized these public datasets. The classification accuracy is calculated by actual labels that are equal to the predicted label divided by total corpus size in test data.

- For Amazon dataset we used review text and summary as features to train the model and for IMDB dataset we used ratings and review text as the feature. Since the Amazon dataset was huge, so for handling dimensionality problem we used Chi2 (Chi- squared) as the feature selection process and selected 50000 features with highest term count to train our model.

4. ANALYSIS AND COMPARISON

In this section, experimental results are explained. During the experiments 5-fold cross validation was applied and performance evaluation parameters were calculated. When the datasets were crawled from the corresponding sites it was unbalanced, and after some preprocessing steps, the distribution of classes (positive and negative) became balanced. For IMDB dataset, we set different thresholds for the movie ratings feature in the range of 2 to 8 to check at which threshold the model performs best and it has been found that the best performance was achieved on threshold values in the range between 4 to 6 as shown in Table 1. For amazon dataset best result was found at threshold value 3. By using threshold mechanism, terms which do not appear frequently in the text are discarded and thereby can improve the overall performance [22]. Although we used unigram (bag of words) for our experiments, this representation can be used for any n-gram (bi-gram, tri-gram etc.).

According to experimental results, the best performance was achieved by Linear SVM based classification model for both of the datasets. Two other classifiers Logistic Regression and Stochastic Gradient Descent also produce almost similar accuracy, precision and also in other measurements. The comparative result of different classifiers are shown in Table 2.

Table 1. Threshold vs. precision of different classifiers (IMDB dataset)

Classifier/Threshold	2	3	4	5	6	7	8
MNB	0.70704	0.6193	0.79402	0.79402	0.79402	0.41289	0.29372
BNB	0.82807	0.81458	0.78636	0.78636	0.78636	0.65726	0.39532
LR	0.84488	0.82244	0.83826	0.83826	0.83826	0.73367	0.54575
Linear SVM @ c=0.25	0.85209	0.82932	0.84354	0.84354	0.84354	0.73491	0.55683
SGD	0.85752	0.82771	0.84416	0.84251	0.84133	0.73467	0.53987
RF	0.70746	0.64088	0.77622	0.77716	0.77794	0.434622	0.29372

The best performance was achieved by our model was using the Linear SVM with parameter selection with selection parameter value c= 0.25 and which was 88.63% for IMDB dataset and 92.18% for Amazon dataset. In SVM approach parameters were optimized during our experiments. As all the experiments were performed using Python 3 and therefore, the results are verifiable and repeatable. We empirically observed that Linear SVM approach has a big potential to improve the performance of sentiment classification model.

Fig. 2 illustrates that the classification accuracy of all the different machine learning algorithms used in our experiment at different threshold values for IMDB dataset. Fig. 3 illustrates the receiver operating characteristics (ROC) curve of different classifiers used in our experiment. An ROC curve is a graph showing the performance of a classifier at different classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (TPR) is also called recall and is defined as:

TPR=TP/(TP+FN)

False Positive Rate (FPR) is called precision and is defined as:

FPR=FP/(FP+TN)

An ROC curve plots TPR vs. FPR at different classification thresholds. The true-positive rate is known as sensitivity and the true-negative rate is known as specificity. An ROC curve shows the tradeoff between sensitivity and specificity. The curves which are closer to the left-hand border and the top border of the ROC space, indicates the more accuracy of that classifiers.

Table 2. Performance analysis of different classifiers

Classifier	IMDB dataset (at t=5)	Amazon dataset
MNB	Matthew Correlation coeff: 0.66998	Matthew Correlation coeff: -0.80846
	Average Precision: 0.79402	Average Precision:0.86854
	F1 Score: 0.81860	F1 Score: 0.90551
	Accuracy: 83.16%	Accuracy: 90.42%
BNB	Matthew Correlation coeff: 0.66231	Matthew Correlation coeff: 0.77882
	Average Precision: 0.78636	Average Precision:0.84496
	F1 Score: 0.81910	F1 Score: 0.84496
	Accuracy: 82.91%	Accuracy: 88.93%
LR	Matthew Correlation coeff: 0.76632	Matthew Correlation coeff: 0.88910
	Average Precision: 0.83826	Average Precision:0.88910
	F1 Score: 0.88321	F1 Score: 0.92326
	Accuracy: **88.32%**	Accuracy: **92.18%**
Linear SVM with parameter selection	Matthew Correlation coeff: 0.77258	Matthew Correlation coeff: 0.84364
	Average Precision: 0.84354	Average Precision: 0.88930
	F1 Score: 0.88582	F1 Score: 0.92332
	Accuracy: **88.63% at c=0.25**	Accuracy: **92.18% @ c=2**
SGD	Matthew Correlation coeff: 0.76891	Matthew Correlation coeff: 0.84231
	Average Precision: 0.84251	Average Precision:0.89080
	F1 Score: 0.88341	F1 Score: 0.922165
	Accuracy: **88.44%**	Accuracy: **92.11%**
RF	Matthew Correlation coeff: 0.67370	Matthew Correlation coeff: 0.67079
	Average Precision: 0.77716	Average Precision: 0.77732
	F1 Score: 0.83989	F1 Score: 0.84197
	Accuracy: 83.66%	Accuracy: 83.5%

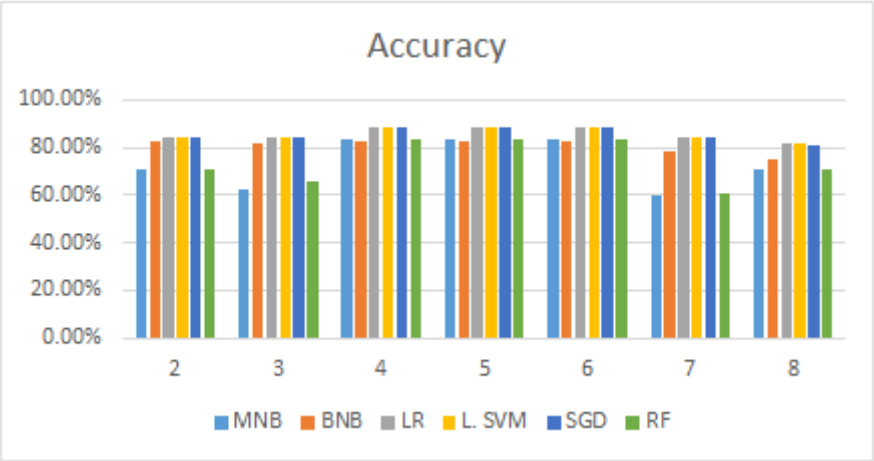
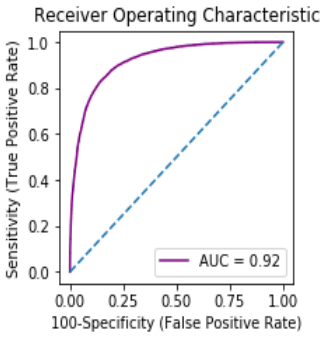
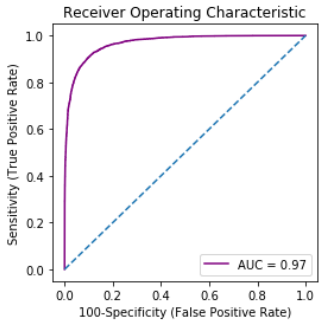


Fig 2: Accuracy of classifiers at different threshold values (IMDB dataset)

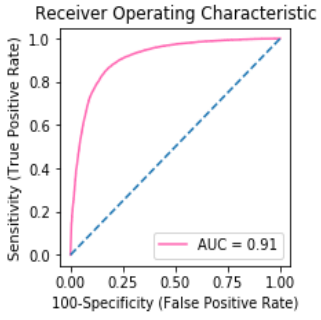
Fig 4 indicates that the accuracy of Logistic Regression, Linear SVM and Stochastic Gradient Descent are almost similar. However, Linear SVM has achieved 0.19 to 0.31% higher average classification accuracy in comparison with the other two, though the difference is not statistically significant. Also it has been found that for linear SVM highest accuracy and precision value was found at c=0.25 and c=2 for the IMDB and Amazon dataset respectively. All other classifier has produce less accuracy than SVM, so inference can be made on that Linear SVM is more stable and less distributed among all other machine learning algorithms for sentiment classification from text review.



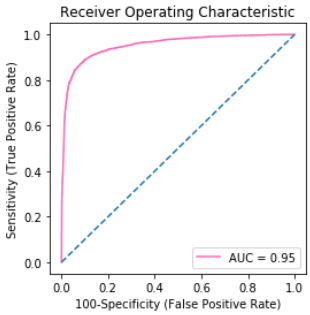
(a) MNB (IMDB dataset)



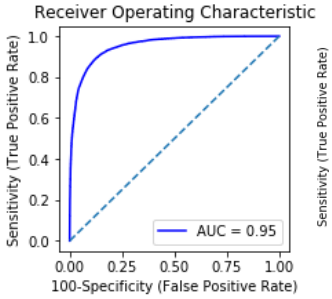
(b) MNB (Amazon dataset)



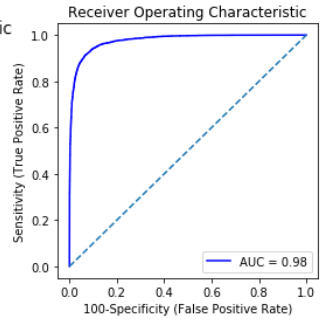
(c) BNB (IMDB dataset)



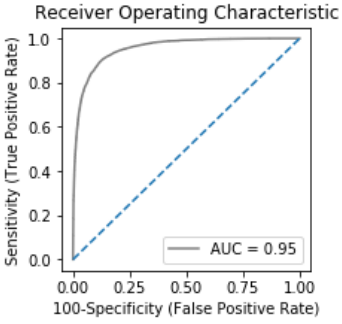
(d) BNB (Amazon dataset)



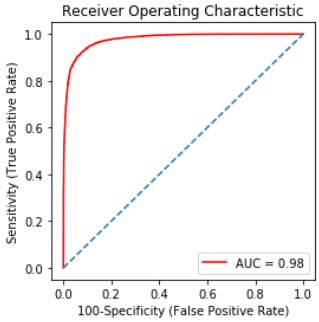
(e) LR (IMDB dataset)



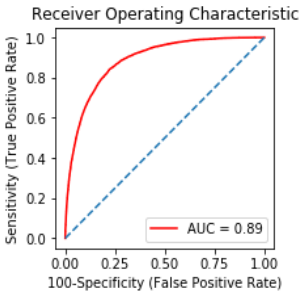
(f) LR (Amazon dataset)



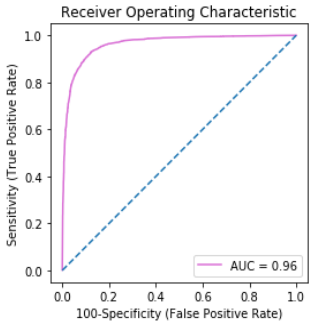
(g) Linear SVM (IMDB dataset)



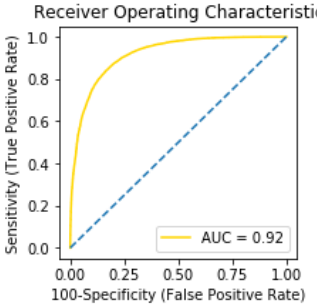
(h) Linear SVM (Amazon dataset)



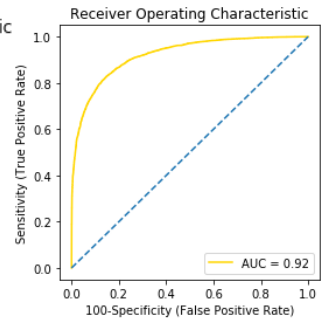
(i) SGD (IMDB dataset)



(j) SGD (Amazon dataset)



(k) RF (IMDB dataset)



(l) RF (Amazon dataset)

Fig. 3: (a)-(l) shows ROC curve of different classifiers

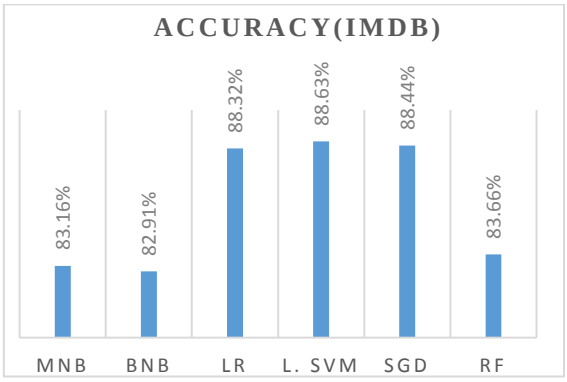
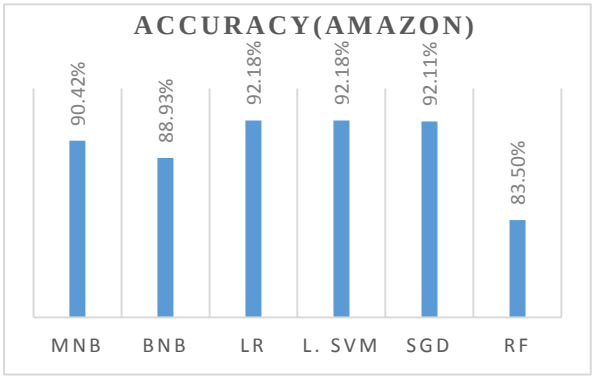


Fig. 4: (a) Accuracy of different classifiers (IMDB dataset)



(b) Accuracy of different classifiers (Amazon dataset)

Though supervised machine learning techniques possesses relatively better performance than unsupervised lexicon based methods in most of the cases, however the main drawback of supervised method is that, it require a large amount of labeled training data that are sometimes very expensive and difficult to collect. Most domains usually have lack of labeled training data and in that case unsupervised methods are very useful. The another limitation of supervised learning is that it

generally requires large expert annotated training corpora to be created from scratch, specifically for the application at hand, and may fail when training data are insufficient.

5. CONCLUSION AND FUTURE WORK

With the fast growth of internet and web technologies the social media serves as a platform to express and share people's feelings, opinions, and comments freely. This rapid growth makes social network as a storage of huge number of reviews about products, services, and solutions. These huge data source not only reflect the changed habits of customers, but also carry information about the brand-customer relationship significantly. Negative or positive experiences spread very quickly by using social platforms such as facebook, orkut or twitter. So it is very much essential for companies, large organizations, policy makers and other key concerns to investigate their big data and steer up the strategies based on the observed findings. Useful information can be discovered by analyzing sentiments from the available user reviews. In this study, multiple machine learning algorithms were investigated to compare their performance for sentiment classification from text reviews. According to experimental results, Linear SVM approach is much better for sentiment classification. This assumption is made after a huge number of experiments by using different classifiers and combinations with two different review datasets. For our work we have focused on specific attribute, but there still some alternative scopes in data pre-processing and attribute selection process that we have plan to do in our future work. Moreover, datasets from various domains like financial, political and social networks can be considered to observe how the accuracy varies according to the variations of dataset. Parameter optimization can be performed by using genetic algorithms and semantics analysis might also improve the performance which is yet to apply as a future work. Neutral messages as well as emoticon features can also be taken into account to improve the overall perfection of the model.

6. REFERENCES

[1] C. Akkaya, J. Wiebe, and R. Mihalcea, "Subjectivity word sense disambiguation", in Proc. Of Conf. Empirical Methods Natural Language Processing, Association Computer Linguistic, vol. 1, pp. 190–199, 2009.

[2] E. Ahmed, M. A. U. Sazzad, M. T. Islam, M. Azad, S. Islam, and M. H. Ali, "Challenges, comparative analysis and a proposed methodology to predict sentiment from movie reviews using machine learning", in International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), pp. 86–91,2017.

[3] https://www.kaggle.com/iarunava/imdb-movie-reviews-dataset[Accessed: 04-Aug-2018].

[4] http://jmcauley.ucsd.edu/data/amazon/ [Accessed: 05-Aug-2018].

[5] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends® in Information Retrieval, vol. 2, no. 1–2, pp. 1–135, 2008.http://dx.doi.org/10.1561/1500000011

[6] B. Liu, "Sentiment Analysis and Opinion Mining", Synth. Lect. Hum. Lang. Technol., vol. 5, no. 1, pp. 1–167, May 2012.

[7] E. Cambria, B. Schuller, Y. Xia and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis", IEEE Intelligent Systems, vol. 28, no. 2, pp. 15–21, 2013.

[8] R. Feldman, "Techniques and applications for sentiment analysis", Commun. ACM, vol. 56, no. 4, p. 82, 2013.

[9] A. Montoyo, P. Martínez-Barco, and A. Balahur, "Subjectivity and sentiment analysis: An overview of the current state of the art and envisaged developments", Decision Support System, vol. 53, no. 4, pp. 675–679, 2012.

[10] M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web", Data Mining and Knowledge Discovery, vol. 24, no. 3, pp. 478–514, 2012.

[11] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Text genre detection using common word frequencies", in Proc. of 18th Conference on Computer Linguistics, vol. 2, p. 808, 2000.

[12] V. Hatzivassiloglou and J. M. Wiebe, "Effects of adjective orientation and gradability on sentence subjectivity", in Proceedings of the 18th conference on Computational linguistics,vol. 1, pp. 299–305, 2000.

[13] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in Proceedings of the 35th annual meeting on Association for Computational Linguistics, pp. 174–181, 1997.

[14] P. D. Turney and M. L. Littman, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", in Proceedings of the Association for Computational Linguistics 40th Anniversary Meeting. Association for Computational Linguistics, New Brunswick, N.J, 2002

[15] G. Grefenstette, "Sextant: Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis", in Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, pp. 324–326, 1992.

[16] A. Huettner and P. Subasic, "Fuzzy Typing for Document Management", in Tutorial Abstracts and Demonstration Notes (ACL 2000 Companion Volume), pp. 26–27, 2000.

[17] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification",Journal of Machine Learning Research, vol. 2, pp. 45–66, 2001.

[18] P. D. Turney, "Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews", in Proc. 40th Annual Meeting of Association Computer Linguistics,pp. 417–424, 2002.

[19] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features", in Proceedings of 10th European Conference on Machine Learning, pp. 137–142, 1998.

[20] S. Dumais, "Using SVMs for Text Categorization," IEEE Intelligent Systems Magazine, Trends and Controversies, pp. 18–28, 1998.

[21] M. Seddon, "Natural Language Processing with Apache Spark ML and Amazon Reviews," [Online] (2015). [Cited: August 10, 2018.] https://mike.seddon.ca/natural-language-processing-with-apache-spark-ml-and-amazon-reviews-part-1/.2015.

[22] A. Mountassir, H. Benbrahim, and I. Berrada, "An empirical study to address the problem of unbalanced data sets in sentiment classification", in Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, pp. 3298–3303, 2012.