Supervised Ensemble Machine Learning Aided Performance Evaluation of Sentiment Classification

Sheikh Shah Mohammad Motiur Rahman\textsuperscript{1, a}, Md. Habibur Rahman\textsuperscript{1, 2, b}, Kaushik Sarker\textsuperscript{1, c}, Md. Samadur Rahman\textsuperscript{3, d}, Nazmul Ahsan\textsuperscript{3, e} and M. Mesbahuddin Sarker\textsuperscript{3, f}

\textsuperscript{1}Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
\textsuperscript{2}Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail, Bangladesh
\textsuperscript{3}Institute of Information Technology, Jahangirnagar University, Dhaka, Bangladesh

\textsuperscript{a}motiur.swe@diu.edu.bd; \textsuperscript{b}mdhabibur.r.bd@iee.org; \textsuperscript{c}kaushik.swe@daffodilvarsity.edu.bd; \textsuperscript{d}engrsamadcse@gmail.com; \textsuperscript{e}ahsan171118@gmail.com, \textsuperscript{f}sarker@juniv.edu

Abstract. Text vectorization, features extraction and machine learning algorithms play a vital role to the field of sentiment classification. Accuracy of sentiment classification varies depending on various machine learning approaches, vectorization models and features extraction methods. This paper represents multiple ways of evaluations with the necessary steps needed to achieve highest accuracy for classifying the sentiment of reviews. We apply two n-gram vectorization models - Unigram and Bigram individually. Later on, we also apply features extraction method TF-IDF with Unigram and Bigram respectively. Five ensemble machine learning algorithms namely Random Forest (RF), Extra Tree (ET), Bagging Classifier (BC), Ada Boost (ADA) and Gradient Boost (GB) are used here. The key findings in this study is to determine which combination of vectorization models (Bigram, Unigram) along with feature extraction method (TF-IDF) and ensemble classifier gives the better performance of sentiment classification.

1. Introduction

Sentiment Analysis is the process to identify the user centric visualization based on user feedbacks. The main concept of sentiment analysis is to analyze the written text by the user and to extract their feedbacks [1]. Sentiment Analysis is used in different areas of healthcare, politics and movie industries [2]. In the industry of healthcare, it is used to find out the quality of patient care at any hospital or health care centre [2]. In the area of politics, the analysis of public’s sentiment towards candidate of political parties [3], to predict the results of election [4]. In the industry of movie, it is used to find out the reaction of people who have watched the movie from their opinions and posted reviews [5]. To perform the analysis, researchers are using various machine learning algorithms which carry out an optimal way to find out the result from people’s sentiment. Two types of machine learning algorithm are used in sentiment analysis, one is supervised and another is unsupervised. In supervised learning, the dataset is labeled during preprocessing or during analysis then trained the machine to obtain a meaningful output that helps to make the decision [6]-[7]. In unsupervised learning, it’s getting harder...
to obtain reasonable result because of unlabeled data. That means unsupervised learning not provide any labeled data to machine. Various clustering methods are used to solve this issue [6]. Again the robustness of supervised learning can be enhance by applying the averaging, boosting and stacking techniques. These are named as ensemble machine learning techniques [8]-[9]. Thus, ensemble machine learning techniques are also known as supervised learning techniques. In this study ensemble machine learning techniques are used to perform classification of sentiment based on labeled data. Mainly three levels are used to investigate the sentiment classification such as aspect level, document level and sentence level [10]. The whole presented expressions and the references aspect within a document is focused on aspect level classification. In document level classification, each and every single document contain either a positive or negative reviews. Whether a sentence is positive or negative can be found out in sentence level classification [10].

This study focused on the aspect of document level classification. The structure of this paper is as follows. Section 2 is provided with the Literature Review based on the sentiment analysis research activities held recently. Section 3 is described the research methodology. Section 4 evaluates the performance to show that results meet up with design and objectives. Section 5 concludes the paper.

2. Literature Review
Zhang et al. [11] have considered word2vec and SVMperf method. They have found 87.10% and 90.30% classification accuracy for word2vec and SVMperf respectively. They use Chinese comment on clothing products dataset and the feature selection methods lexicon-based and part-of-speech have used. Martin et al. [12]. have used voting and stacking concept with machine learning algorithms. Implementing the document classification and sentiment classification they use Naïve Bayes, Support Vector Machine, Bayesian Logistic Regression (BLR) and C4.5. Using the MuchoCine(MC) and Spanish Corpus dataset they found the accuracy 88.5%, 88.31%, 88.27% and 87.66% respectively for the above algorithms. Xia et al. [13] have found 82.7%, 86.1% and 84.85% classification accuracy for Naïve Bayes, Support Vector Machine and Maximum Entropy respectively. The mostly use Cornell movie review dataset, document and multi-domain sentiment dataset. For selecting features set, they have considered part-of-speech(POS) and the word-relation based feature sets. Matsumoto et al. [14] have worked with document level sentiment analysis and for sentiment classification they use support vector machine and apply on Internet Movie Database (IMDb) and Polarity Dataset. They have found the classification accuracy 83.7% for Unigram, 80.4% for Bigram and 84.6% for Unigram + Bigram using only support vector machine learning algorithm. Tripathy et al. [6] have considered document label classification with CountVectorizer for vectorization along with TF-IDF to transform text to numeric values. The have used Naive Bayes (NB), Support Vector Machine (SVM), Random Forest and Linear Discriminant Analysis (LDA) machine learning algorithm for sentiment classification. They found 89.5%, 94.0%, 95.0% and 92.0% accuracy form Naive Bayes (NB),Support Vector Machine (SVM), Random Forest, and Linear Discriminant Analysis (LDA) respectively. They use aclIMDb [15] and polarity movie [16] review dataset. However highest accuracy into 100% is not found for any classification algorithm yet. This study try to found the highest accuracy for classifying sentiment of reviews. Random Forest (RF), Extra Tree (ET), Bagging Classifier (BC), Ada Boost (ADA) and Gradient Boost (GB) machine learning algorithms are used to test the accuracy level.

3. Methodology
The process of sentiment classification are two types: binary sentiment classification and multiclass sentiment classification [4]. Each document doc, in DOC, where DOC = {doc1, doc2, doc3, ..., docn} is classified in binary classification. It is labeled as S, where, S is a predefined set of review category like S = {1,0} = {Positive, Negative}. Each document doc is classified in multi class sentiment analysis as a level in S* where, S* = {Strong positive, positive, neutral, negative, strong negative} [6]. Binary sentiment classification is considered in this study. The overall system is described in figure 1.
3.1 Dataset Used
The dataset used for our experiment is “Large Movie Review Dataset” collected by Andrew Maas [17] [18]. This dataset consists of a collection of 50,000 movie reviews from IMDb [15]. The movies are covered by the category comedy, funny, action, science fiction, romantic, horror, fantasy and adventure. At most 30 reviews are considered as per movie. The movie reviews are classified into equal number of positive and negative reviews. Neutral reviews are ignored. Based on the user rating out of 10, less than or equal 4 is classified as a Negative review and greater than or equal 7 is classified as positive reviews.

3.2 Vector Creation and Feature extraction
Generate a vectorized documents by n-Grams of tokens (unigram and bigram) methods from each textual reviews. A term n-Gram is defined as a series of consecutive tokens of length n. Feature extraction is an important method to determine the robust key features from large number reviews [19] [20]. Feature extraction methods TF-IDF is used in this research which has implemented on both of unigram and bigram methods [21].

3.3 Evaluation with Ensemble Machine Learning
After that, the performance of classifying the sentiment of reviews from vectorized documents which are generated using the Unigram, Bigram, Unigram + TF-IDF and Bigram + TF-IDF models on dataset is being evaluated. Five ensemble machine learning approaches (Bagging Classifier, Random Forest, Extra Tree, Ada Boosting, Gradient Tree Boosting) are applied to classify the sentiment of reviews and to calculate the accuracy of classification [22].

4. Performance Evaluation and Result Analysis
This section analyzes the accuracy level and the overall performance of the proposed model. Bigram and Unigram vectorization models and a well established feature extraction methods (TF-IDF) have been implemented and evaluated in this study. Here ensemble machine learning algorithms are considered [23]. Several feature extraction methods are applied for evaluating the performance [24].

4.1 Evaluation Parameter
To evaluate the performance of the classification of sentiment of reviews, there are some metrics are used which are as outlined in Table 1.
Table 1. Evaluation Parameter for Measuring the Accuracy [5] [6].

| Evaluation Parameter | Description | Formula/Visualization |
|----------------------|-------------|-----------------------|
| Precision            | It measures the exact value of the result provided by the classifier. | Positive Precision = TP/(TP + FP) Negative Precision = TN/(TN+FN) |
| Recall               | It measures the thoroughness of the result provided by classifier. | Positive Recall = TP/(TP + FN) Negative Recall = TN/(TN+FP) |
| F1-Score             | It defines the harmonic mean of exactness and completeness. | F1-Score = (2*Precision* Recall) / (Precision + Recall) |
| Accuracy             | It defines the ratio of correctly classified data to total number of data. | Accuracy = (TP + TN) / (TP + TN + FP + FN) |

TP : True Positive represents the reviews are positive and also classifier classified as positive; FP: False Positive represent the reviews are positive but classifier classified as negative; TN: True Negative represents the reviews are negative and also classifier classified as negative; FN: False Negative represent the reviews are negative but classifier classified as negative; [5]

Figure 2 depicts the results obtained from the Unigram and Bigram vectorization model with five ensemble machine learning techniques. The blue and maroon color represents the accuracy of Unigram and Bigram vectorization model respectively. We obtain 99.404%, 100%, 98.664%, 80.426% and 81.722% accuracy by using Random Forest (RF), Extra Tree (ET), Bagging Classifier (BC), ADA Boost (ADA) and Gradient Boost (GB) respectively for Unigram vectorization model. On the other hand, the accuracy for Bigram vectorization model obtained from Random Forest(RF), Extra Tree(ET), Bagging Classifier(BC), ADA Boost (ADA) and Gradient Boost(GB) classifiers are 99.404%, 100%, 98.754%, 80.474% and 81.76% respectively represented in figure 2.

Figure 2. Accuracy for Unigram and Bigram Vectorization Model using RF, ET, BC, ADA and GB Classifiers
Figure 3. Accuracy for Unigram+TF-IDF and Bigram+TF-IDF Vectorization Model using RF, ET, BC, ADA and GB Classifiers

The results obtained from the Unigram + TF-IDF and Bigram + TF-IDF vectorization model with five ensemble machine learning techniques represented in figure 3. The accuracy of Unigram + TF-IDF and Bigram + TF-IDF vectorization model represented respectively by blue and maroon color. The accuracy of classification obtained from Random Forest(RF), Extra Tree(ET), Bagging Classifier(BC), ADA Boost (ADA) and Gradient Boost(GB) classifiers are 99.38%, 100%, 98.892%, 80.814% and 82.024% respectively using Unigram + TF-IDF technique. On the other hand, we observe that Random Forest(RF), Extra Tree(ET), Bagging Classifier(BC), ADA Boost (ADA) and Gradient Boost(GB) algorithms provide the accuracy of classification 99.47%, 100%, 98.914%, 80.516% and 82.258% respectively using Bigram + TF-IDF technique. All of the experimental result with the results of evaluation parameters are comparatively shown at table 2.

Table 2. Comparison among the combination of Experimental Results.

| Classifier + N-Gram | Precision | Recall | F1-Score | Accuracy |
|---------------------|-----------|--------|----------|----------|
| RF + Uni            | 0.99      | 0.99   | 0.99     | 99.4%    |
| ET + Uni            | 1.00      | 1.00   | 1.00     | 100%     |
| BC + Uni            | 0.99      | 0.99   | 0.99     | 98.7%    |
| ADA + Uni           | 0.81      | 0.80   | 0.80     | 80.4%    |
| GB + Uni            | 0.82      | 0.82   | 0.82     | 81.7%    |

| Classifier + N-Gram | Precision | Recall | F1-Score | Accuracy |
|---------------------|-----------|--------|----------|----------|
| RF + Bi             | 0.99      | 0.99   | 0.99     | 99.4%    |
| ET + Bi             | 1.00      | 1.00   | 1.00     | 100%     |
| BC + Bi             | 0.99      | 0.99   | 0.99     | 98.8%    |
| ADA + Bi            | 0.81      | 0.80   | 0.80     | 80.4%    |
| GB + Bi             | 0.82      | 0.82   | 0.82     | 81.8%    |
## 5. Conclusion

The combination of text vectorization models without and with feature extraction method TF-IDF are broadly experimented with five ensemble machine learning algorithms. Table 2 represent evaluation parameters and accuracy obtained and Figure 2 and 3 represent the accuracy obtained from Unigram, Bigram, Unigram + TF-IDF and Bigram + TF-IDF respectively after classification using five Ensemble Machine Learning classifiers. We found the maximum classifier can provide highest accuracy from the combination of Bigram + TF-IDF. And Extra Tree provide 100% accuracy for all combinations, ADA Boost for Unigram + TF-IDF and rest three classifiers Random Forest, Bagging Classifier and Gradient Boost algorithm for Bigram + TF-IDF combination.

### References

[1]. Arif, M. H., Li, J., Iqbal, M., & Liu, K. Sentiment analysis and spam detection in short informal text using learning classifier systems. Soft Computing, 1-11, (2017).

[2]. Elnagar, A., Khalifa, Y. S., & Einea, A. Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications. In Intelligent Natural Language Processing: Trends and Applications (pp. 35-52). Springer, Cham, (2018).

[3]. Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the Association for Computational Linguistics System Demonstrations. pp. 115-120, (2012).

[4]. Liu, B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, vol. 5(1), pp. 1-167. (2012).

[5]. Sorostinean, M., Sana, K., Mohamed, M. O. H. A. M. E. D., & Targhi, A. Sentiment Analysis on Movie Reviews, (2017)

[6]. Tripathy, A., & Rath, S. K. Classification of sentiment of reviews using supervised machine learning techniques. International Journal of Rough Sets and Data Analysis (IJRSDA), 4(1), 56-74, (2017).

[7]. Gautam, G., & Yadav, D. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In seventh international conference on Contemporary computing (IC3), pp. 437-442. IEEE, (2014).

[8]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. & Vanderplas, J. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830., (2011).

[9]. Ensemble methods, http://scikit-learn.org/stable/modules/ensemble.html.

[10]. Feldman, R. Techniques and applications for sentiment analysis. Communications of the ACM, vol. 56(4), pp. 82-89, (2013).

[11]. Zhang, D., Xu, H., Su, Z., & Xu, Y. Chinese comments sentiment classification based on

| Classifier +N-gram | Precision | Recall | F1-Score | Accuracy |
|--------------------|-----------|--------|----------|----------|
| RF+Uni+TF-IDF      | 0.99      | 0.99   | 0.99     | 99.38%   |
| ET+Uni+TF-IDF      | 1.00      | 1.00   | 1.00     | 100%     |
| BC+Uni+TF-IDF      | 0.99      | 0.99   | 0.99     | 98.9%    |
| ADA+Uni+TF-IDF     | 0.81      | 0.81   | 0.81     | 80.8%    |
| GB+Uni+TF-IDF      | 0.82      | 0.82   | 0.82     | 82.0%    |

| Classifier+N-gram | Precision | Recall | F1-Score | Accuracy |
|--------------------|-----------|--------|----------|----------|
| RF+Bi+TF-IDF       | 0.99      | 0.99   | 0.99     | 99.5%    |
| ET+Bi+TF-IDF       | 1.00      | 1.00   | 1.00     | 100%     |
| BC + Bi+ TF-IDF    | 0.99      | 0.99   | 0.99     | 98.9%    |
| ADA+Bi+TF-IDF      | 0.81      | 0.81   | 0.80     | 80.5%    |
| GB+Bi+TF-IDF       | 0.83      | 0.82   | 0.82     | 82.3%    |
word2vec and SVMperf. Expert Systems with Applications, vol. 42(4), pp. 1857-1863, (2015).
[12]. Martín-Valdivia, M. T., Martínez-Cámaro, E., Perea-Ortega, J. M., & Ureña-López, L. A. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications, vol. 40(10), pp. 3934-3942, (2013).
[13]. Xia, R., Zong, C., & Li, S. Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, vol. 181(6), pp. 1138-1152, (2011).
[14]. Matsumoto, S., Takamura, H., & Okumura, M. Sentiment classification using word subsequences and dependency sub-trees. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 301-311, Springer, Berlin, Heidelberg, (2005).
[15]. Large Movie Review Dataset, http://ai.stanford.edu/~amaas/data/sentiment/.
[16]. Movie Review Data, http://www.cs.cornell.edu/people/pabo/movie-review-data/.
[17]. Tang, H., Tan, S., & Cheng, X. A survey on sentiment detection of reviews. Expert Systems with Applications, vol. 36(7), pp. 10760-10773, (2009).
[18]. Types of language models, https://nlp.stanford.edu/IR-book/html/htmledition/types-of-language-models-1.html.
[19]. Croft, W. B., Metzler, D., & Strohman, T. Search engines: Information retrieval in practice vol. 283. Reading: Addison-Wesley, (2010).
[20]. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1, No. 1, p. 496). Cambridge: Cambridge university press.
[21]. Tf-idf, http://www.tfidf.com/
[22]. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1 (pp. 142-150). Association for Computational Linguistics, (2011).
[23]. Language Modeling, https://web.stanford.edu/class/cs124/lec/languagemodeling.pdf.
[24]. Ahmad, F. K. Comparative Analysis of Feature Extraction Techniques for Event Detection from News Channels’ Facebook Page. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 9(1-2), pp. 13-17, (2017).