Analysis of Machine Learning Approaches for Opinion Mining of Movie Reviews

[1]Ramandeep Sharma, [2]Sukhjit Singh Sehra, [3]Sumeet Kaur Sehra
[1] Research Scholar, [2] [3] Assistant Professor

***Abstract: -* Opinion Mining plays a vital role in the area of machine learning, data mining, and natural language processing. This paper describes various sentiment analysis techniques, feature extraction processes, and challenges that make the sentiment analysis difficult. The three subtasks are performed for sentiment analysis: text pre-processing, feature extraction and classification. In this paper, three machine learning techniques (Naïve Bayes, Support Vector Machine (SVM) and Decision Tree) are used to classify the movie reviews. The highest accuracy was achieved with SVM classifier using Movie Reviews when unigrams were used as a feature. The Naïve Bayes and SVM obtained 78.70% accuracy with movie reviews dataset using bigrams as a feature.**

***Index terms:* Classification, Machine Learning Techniques, Movie Reviews, Sentiment Analysis,.**

I. INTRODUCTION

The rapid proliferation of Web 2.0 has made it possible for people to express their opinions over the Internet[1]. It gathers a huge volume of opinionative data in the last few decades. Users express their views, thoughts or feelings on Social media or on review sites about product or movies in a very convenient way. Expressed views or thoughts are their sentiments and opinions about the particular topic. The reviews contain knowledgeable and important information for any decision-making process. Due to its credibility, many researchers are motivated to work on sentiment analysis or opinion mining. Earlier, Companies preferred questionnaires to take the decision about their products[2]. On the other hand, many people are not interested in filling questionnaires, and it is also a time-consuming process. Nowadays, users express their opinions about the products on reviews sites easily. But the content available on the web is too vast which is difficult for a user to analyze and classify the sentiment of the reviews[3]. To solve this problem, different sentiment analysis techniques are used. The main aim of opinion mining is to determine the emotions, opinions, and attitude in short texts, documents, sentences from blogs, reviews, and news. In Opinion Mining, the semi-structured and unstructured information is processed and labels the sentiments into different categories like positive, negative or neutral. It identifies whether a given text expresses the author's opinions (subjective) or expresses the factual information (objective). Sentiment analysis is used to find the emotions and opinions about the particular topic. For example, companies can use sentiment analysis to find whether people like their products and services or not. They may also want to know if people give positive or negative reviews about their products and would people prefer their products for use. Manufacturing companies analyze the sentiments to collect the or to investigate customer satisfaction. Political parties can take help from sentiment analysis to judge whether public support their decisions and party or not. This information can be obtained from social sites because the users of social media post daily about what they like or dislike.Various Machine learning Techniques are Naïve Bayes, Support Vector Machine and Decision tree etc.

II. LITERATURE SURVEY

Neethu and Rajasree[3] used different machine learning techniques and new feature vector to classify the sentiment of the tweets about electronic products. They used a Stanford POS tagger for tagging the part of speech in the sentence and performed sentence-level sentiment analysis. Different classifiers performance was measured and found that Maximum Entropy Classifier, SVM, and Ensemble classifiers achieved an accuracy of 90% and Naïve Bayes achieved 89.5%.

Mukwazvure and Supreethi[4] proposed an approach to detect the sentiments of news corpus from the Politics, Technology, and Business sections and use opinion lexicon to identify their sentiments. They used unigram feature for SVM and K-Nearest Neighbor (k-NN) on news comments and found that SVM performed better than k-NN. In k-NN, a larger value of k attained better accuracy than lower value of k and a small dataset gave poor classifier performance. The accuracy of Technology section was

73.36% with SVM classifier and 74.24% with k-NN classifier when k=16. feedback about their products and issues related to the products. Marketers may be interested to monitor the public opinions about the products and company, Bermingham and Smeaton[5] considered two supervised(SVM and Multinomial Naive Bayes (MNB)) and one unsupervised classifier (SentiWordNet) for four datasets and found that SVM attained 87.9% accuracy on movie reviews with Unigram+Bigram+Trigram and Unigram+POS n-gram (n=1) feature sets.

Bhoir and Kolte[6] performed subjectivity analysis and summarize the movie reviews at aspect level. To find the feature-opinion pair and various aspects of movie reviews, they considered the subjectivity of the sentences and rule-based system. Naïve Bayes and SentiWordNet (SWN) were used for classification and found that naïve Bayes(71.42%) performed better than SWN(53.33%).

Jin et al.[7] used the naïve bayes classifier to improve the accuracy of Movie Reviews using improved Information Gain and decreased the impact of low-frequency words.

Bhadane et al.[8] examined two-step method i.e. aspect classification followed by polarity classification for the classification of natural language text. They implemented various techniques for aspect identification and polarity identification. To do so, a model for aspect classification and a model per aspect for polarity classification were built and achieved 78.05% accuracy for polarity classification

Whitehead and Yaeger[9] considered the ensemble methods (bagging, random subspace, boosting, and bagging random subspaces) for increasing the classification accuracy of the restaurant reviews collected by themselves and one used by Snyder and Barzilay[10]; SVM was used as a base model for ensemble methods and unigrams were chosen as feature. The random subspace and bagging random subspaces ensembles achieved better accuracy than the bagging ensemble. If high classification accuracy is required and there is no time and computational resources constraints, then bagging subspace model considers being the best choice.

Haddi et al.[11] examined the role of text pre-processing in investigating the sentiments of online movie reviews and found that selecting appropriate features and representation, sentiment analysis accuracies can be improved using SVM. They used different pre-processing methods and chi-squared methods to reduce the noise and remove irrelevant features. They used the non-preprocessed and preprocessed data for classification with the features matrices(TF-IDF, FF, FP) and compared the obtained results with Pang et al.[12] for TF-IDF and FF matrices. The accuracy of FF, FP, and TF-IDF matrix was 76.33%, 82.33%, and 78.33% respectively without pre-processing the data of movie reviews and the accuracy of FF and FP matrix obtained in Pang et al.[12] was 72.8% and 82.7% respectively but Pang et al.[12] did not use TF-IDF. After applying chi-squared feature selection, the obtained accuracy was 93%, 92.3%, and 90% in FP, TF-IDF, and FF matrix.

HLTCOE[13] presented Sentiment Analysis in Twitter which incorporates two tasks: A, expression level, and B, message level. They used twitter corpus which consists of tweets and SMS messages for training and testing purpose and for testing purpose respectively. Task A gave better results than task B. The various teams submitted their results for twitter test set- constrained and unconstrained systems. The average F1-measure evaluated by NRC-Canada and GU-MLT-LT team for twitter test set and SMS test set was 88.93% and 88.37% respectively. NRC-Canada obtained average F1-measure for twitter test set and SMS test set was 69.02% and 68.46% respectively.

Basari et al.[14] used Hybrid Approach of Support Vector Machine and Particle Swarm Optimization for Sentiment classification of movie reviews. They used case normalized, tokenized, stemmed and generated n-grams as features and tf and tf-idf as weighting techniques. An SVM-PSO technique incorporated two machine learning techniques to improve the SVM using PSO. SVM-PSO gave higher accuracy and precision than SVM alone.

Appel et al.[15] presented the hybrid method which uses NLP techniques, a sentiment lexicon with SentiWordNet and fuzzy sets to classify the sentiments at the sentence level. Hybrid Standard Classification and Hybrid Advanced Classification is applied to three data sets and 88.02% accuracy achieved with a hybrid approach using twitter dataset which is more than the accuracy obtained with naïve Bayes and Maximum Entropy.

Lin et al.[16] developed particle swarm optimization (PSO) for determining the parameters and selecting the features of SVM. The PSO+SVM approach attained high accuracy with appropriate parameters and subsets. The results of PSO+SVM obtained for public datasets are compared with GA+SVM proposed by Huang et al.[17].

Lin and Yu[18] presented the weighted naïve bayes classifier based on particle swarm optimization and it enhanced the accuracy of naïve bayes classifier. The accuracy rate of particle swarm optimization based weighted naive Bayesian classifier (PSOWNBC) was higher than Naïve bayes classifier but algorithm running time of PSOWNBC was slightly higher than NBC.

Ghag and Shah[19] presented a survey on different techniques used in Sentiment classification and comparison of them on the basis of requirement of the training set, usage of lexicon and language dependency. They have also discussed the challenges of sentiment analysis. Some of the challenges are handling negations, polysemy, slangs and domain generalization, Language Generalization, Feature Matrix Construction, Hidden Sentiments Identification.

Patil and Atique[20] explored the challenges that make the sentiment analysis difficult as compared to traditional text-based analysis. Several challenges are Sarcasm, slangs, Review Author Segmentation, handling negations, and polysemy. These challenges provide opportunities for future research. This survey provided a brief description of recent articles and techniques used. They were also found that lot of work has done on machine learning methods rather than lexicon-based method.

Madhoushi et al.[21] presented the survey to categorize the sentiment analysis techniques without focusing on specific task or level. They found open problems in SA which are still unsolved in this field. Insufficient labeled data is the challenging problem in SA. Very few research articles on SA are present in a language other than English. Research should be done in other languages also and existing techniques are still unable to deal with typical sentences.

Medhat et al.[22] presented algorithms improvement, summarization, and categorization of the articles and techniques with brief details of the algorithms of Sentiment analysis. They investigated various applications and fields related to SA that includes emotion detection, transfer learning, and building resources. They found some open problems in research such as Data Problem and Language problem.

III. METHODOLOGY

The methodology of sentiment analysis is categorized into four subtasks: (1) collecting datasets, (2) performing text pre-processing which incorporates eliminating stop words, filtering numbers, special characters, punctuation, performing stemming and POS tags (adverbs, verbs and adjectives), (3) feature extractions such as unigrams and bigrams. (4)Text Classification, to classify the text using supervised learning classifiers. TF and TF-IDF are used for feature selection and unigrams and bigrams are used as features. We split both the datasets into 70% training set and 30% testing set.

A. Dataset

The twitter dataset and movie reviews dataset are widely used for sentiment analysis. We have used the movie reviews corporus for sentiment classification. We used a set of 1000 movie reviews for classification which contains 500 positive and 500 negative reviews.

B. Text Preprocessing

Preprocessing refers to removing the unnecessary and irrelevant data from the dataset to increase the performance of classification. The following preprocessing steps have been done in our project to enhance the classification:

***Case Converter*: It converts all the terms incorporated in the documents to upper or lower case.

***Stop Word Removal*: Stop words are eliminated from the documents because they do not affect the meaning of the sentence. The English language has some stop words like a, is, on, an, of, the etc. We used the built-in stop word list for removing them.

Stemming: Stemmer stems the terms present in the documents with the stemming libraries. Stemmer used to reduce the feature set and enhance the performance of the classification. We are using Porter stemmer for stemming the words of the dataset.

***Punctuation Erasure*: It removes all punctuation characters of terms present in the documents.

***Number Filter*: It filters all terms incorporated in the documents that consist of numbers, and decimal separators and arithmetic operations.

***N Chars Filter*: It filters all terms present in the documents with less than N characters. We have used N=3 for removing all words whose length is less than 3, except none, no, not.

***Part-of-speech tagging*: POS tagging converts a sentence, in the form of (word, tag). The POS tag indicates whether the word is a verb, adverb, adjective or noun etc.

***Features Selection and Feature Extraction*: Feature selection means to select the attributes in the data that are relevant to the predictive modeling problem. The number of features can be identified with feature identification process and feature weighting scheme is used to select the best feature because some features have less contribution in classification.

***Unigram and Bigrams*: We are using unigrams and bigrams as a feature for all classifiers.

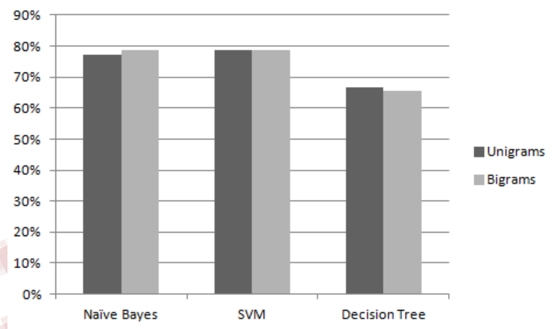
***Bag of Words creation*: From a given documents, a bag of words are created which consists of two columns, one column containing a document and other containing the terms occurring in the document.

Term Frequency (TF): Term Frequency measures how often a term presents in a document. Some low-frequency terms may be ignored in term frequency but in some cases, low-frequency terms have also a great contribution in classification so different term weighting methods are employed. The term frequency is calculated by dividing the number of times term present in a document by the total number of terms present in the corresponding document.

TF-IDF: TF-IDF is the approach we used in this work. It measures how important a term is to a document in a corpus. The importance of TF-IDF increases proportionally to the number of times a term appears in the document.

IV. RESULTS AND DISCUSSION

The various experiments are performed on the movie reviews dataset. In this work, four subtasks are performed which includes data collection, text pre-processing, feature extraction and classification. We used Naïve Bayes, Support Vector Machine, and Decision Tree classifiers for analyzing and classifying the movie reviews using unigrams and bigrams as features with dataset size 1000 and found that SVM classifier obtained the accuracy of 78.70% with movie reviews when unigrams were used as a feature. The Naïve Bayes and SVM obtained 78.70% accuracy with movie reviews dataset using bigrams as a feature. The accuracies of all the classifiers using unigrams and bigrams are shown in figure 1.



REFERENCES

[1] M. Castellanos, U. Dayal, M. Hsu, R. Ghosh, M. Dekhil, Y. Lu, L. Zhang, and M. Schreiman, “LCI: a social channel analysis platform for live customer intelligence,” in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, 2011, pp. 1049–1058.

[2] A. Ortigosa, J. M. Martín, and R. M. Carro, “Sentiment analysis in Facebook and its application to e-learning,” *Comput. Hum. Behav.*, vol. 31, pp. 527–541, Feb. 2014.

[3] M. S. Neethu and R. Rajasree, “Sentiment analysis in twitter using machine learning techniques,” in *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*, 2013, pp. 1–5.

[4] A. Mukwazvure and K. P. Supreethi, “A hybrid approach to sentiment analysis of news comments,” in *Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2015 4th International Conference on*, 2015, pp. 1–6.

[5] A. Bermingham and A. F. Smeaton, “Classifying sentiment in microblogs: is brevity an advantage?,” in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 1833–1836.

[6] P. Bhoir and S. Kolte, “Sentiment analysis of movie reviews using lexicon approach,” in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2015, pp. 1–6.

[7] L. Jin, W. Gong, W. Fu, and H. Wu, “A Text Classifier of English Movie Reviews Based on Information Gain,” 2015, pp. 454–457.

[8] C. Bhadane, H. Dalal, and H. Doshi, “Sentiment Analysis: Measuring Opinions,” *Procedia Comput. Sci.*, vol. 45, pp. 808–814, 2015.

[9] M. Whitehead and L. Yaeger, “Sentiment mining using ensemble classification models,” in *Innovations and advances in computer sciences and engineering*, Springer, 2010, pp. 509–514.

[10] B. Snyder and R. Barzilay, “Multiple Aspect Ranking Using the Good Grief Algorithm.,” in *HLT-NAACL*, 2007, pp. 300–307.

[11] E. Haddi, X. Liu, and Y. Shi, “The Role of Text Pre-processing in Sentiment Analysis,” *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013.

[12] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002, pp. 79–86.

[13] J. HLTCOE, “SemEval-2013 Task 2: Sentiment Analysis in Twitter,” *Atlanta Ga. USA*, p. 312, 2013.

[14] A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, “Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization,” *Procedia Eng.*, vol. 53, pp. 453–462, 2013.

[15] O. Appel, F. Chiclana, J. Carter, and H. Fujita, “A hybrid approach to the sentiment analysis problem at the sentence level,” *Knowl.-Based Syst.*, May 2016.

[16] S.-W. Lin, K.-C. Ying, S.-C. Chen, and Z.-J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," *Expert Syst. Appl.*, vol. 35, no. 4, pp. 1817–1824, Nov. 2008.

[17] C.-L. Huang and C.-J. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," *Expert Syst. Appl.*, vol. 31, no. 2, pp. 231–240, Aug. 2006.

[18] J. Lin and J. Yu, "Weighted naive bayes classification algorithm based on particle swarm optimization," in *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on*, 2011, pp. 444–447.

[19] K. Ghag and K. Shah, "Comparative analysis of the techniques for Sentiment Analysis," in *Advances in Technology and Engineering (ICATE), 2013 International Conference on*, 2013, pp. 1–7.

[20] H. P. Patil and M. Atique, "Sentiment Analysis for Social Media: A Survey," in *Information Science and Security (ICISS), 2015 2nd International Conference on*, 2015, pp. 1–4.

[21] Z. Madhoushi, A. R. Hamdan, and S. Zainudin, "Sentiment analysis techniques in recent works," in *Science and Information Conference (SAI), 2015*, 2015, pp. 288–291.

[22] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014.