International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper Volume-8, Issue-07, 2021 Regular Edition E-ISSN: 2349-7084

Sentiment Analysis on Movie Review Data Using Ensemble Vote Classifier Technique

[1]A. Naresh, [2]P. Venkata Krishna , [3]M Bhavsingh

[1]Department of Computer Science ,Bharathiar University,Tamilnadu, India

[2] Department of Computer Science,Sri Padmavati Mahila Visvavidyalayam,Tirupati, India

[3] Department of Computer Science and Engineering, G.Pullaiah College of Engineering and Technology,Kurnool,AP,India

Email ID: pandu5188@gmail.com , pvk@spmvv.ac.in, bhavsinghit@gmail.com

Available online at: http://www.ijcert.org

Received: 09/06/2021, Revised: 2/07/2021, Accepted: 16/07/2021, Published: 29/07/2021

Abstract: Sentiment analysis is a procedure of investigating sentiments or feelings expressed in the text. It classifies whether the given text is positive or negative or sometimes neutral also, based on the classification level on a given document or sentence. To encourage the customers in better decision making and to find perspectives on others and furthermore helps in choosing the buy with the opinion of different customers. In this article a novel ensemble vote classifier technique with logistic regression, random forest and XGB classifiers is proposed and sentiment analysis on real-time movie review data is analyzed using proposed technique. Experimental results shows improved accuracy with balanced class weight compared to without balanced class weight.

Keywords: — classifiers; Ensemble methods; machine learning; Sentiment analysis.

---

1. Introduction

Sentiment analysis-based research is increasingly used in many fields where data processing is carried on linguistics, natural languages that results in required text [1-2]. Sentiment examination often referred to as opinion mining because its task is to detect polarity classification [3]. Extremity grouping is the errand of naming an assessment record as communicating it is possible that it is certain or negative. The order of extremity assessment recognition should be possible at a few levels, for example, term, expression, and sentence or record level. Methods utilizing the n-gram classifiers or vocabularies ordinarily deal with term level. A standout amongst the most essential errands in opinion examination is distinguishing the semantic introduction of a word [1-3]. Now a day's social media have become the forum for discussion and source of information that contain opinionated text. Such text has become the major hub to mining opinions and analysis of user behavior. It provides insights for product feedback, user interaction and lead generations [4-5].

Generally, sentiment analysis was performed by creating a vocabulary of words with negative and positive polarities and recognizing the client's perspective by separating words in the content with the dictionary [6-7]. Sentiment Analysis is a challenging research problem for which various strategies like machine learning, lexicon-based and hybrid methodologies are proposed by researchers as of now. Supervised and unsupervised learning are the two classes of machine learning method. In Supervised learning, the polarity of the test dataset sentiment based on the trained dataset is predicted. Unsupervised and semi-supervised techniques are proposed when it is not possible to provide prior training dataset or labeled documents or opinions to classify the rest of the items [8-9].

Performing sentiment classification of the source context by using various machine learning techniques with improved performance accuracy is a challenging task. This article focuses on various factors and challenges that effect the performance of the sentiment analysis and it also provides an ensemble technique such as ensemble vote classifier with base classifiers as logistic regression, random forest and XGB classifiers. The designed sentiment analysis model chosen for sentiment classification using variety of classifiers can be binomial if exists target classes as positive

or negative or it can be multinomial if the model output classifier classifies the target with more than two classes.

2. Related Work

In [10], discussed a comparative study among traditional and current optimization models for stock selection using investor views as sentiment data. Genetic algorithms are applied on data and concludes that evaluation-based models give better results than traditional regression models. In the area of stock selection all ML techniques are applicable and the team proven that the proposed model given better ranking for stocks by considering stock data of all months.

In [11], proposed an innovative technique bi-directional long- and short-term memory (BLSTM) for text processing. Here the similarity between words are classified by analyzing two level vector representation of text. Each word is represented as vector using word2vec method and each sentence or document is represented as vector using doc2vec method. The experiments are done on Bengali sentences data and classified in two ways successfully as positive and negative. The accuracy of proposed BLSTM is compared with Machine Learning algorithms as part of experimental results.

In [12], presented an operation ML approaches on yelps database to suggest a brand by evaluating user views on individual items. The authors tested the proposed model using various ML techniques such as naïve bayes, random forest, support vector machine, decision trees, K-nearest neighbor and multilayer perception classifies. Experimental results evidence that multilayer perception classifies given better results on sentiment analysis for the identifies user reviews data.

In [13], the authors utilized ML techniques to evaluate sentiment analysis of user reviews on mental health apps. Based on user reviews the sentiment polarity is assessed and classified the apps as positive and negative. Using the proposed system, the authors done sentiment analysis by collecting the user review data related to 104 mental health apps on google play store. Various categories are defined to classify the review data as negative or positive. Five ML classifiers are applied and for classification and the results have compared each other. The identified negative categories help to improve the effectiveness of mental health apps.

In [14], discussed the limitations of various novel methods available for sentiment classification in cross-domain sentiment analysis. The authors proposed an automated model for sentiment analysis with the combination of supervised machine learning approaches. Categorizing the opinions as negative, positive, or neutral based on expressed user reviews will help the people to build new business models.

In [15], the authors presented a case study related to opinion mining techniques by taking Amazon product reviews. The authors applied lexicon dictionary-based technique with ngrams using three ML algorithms. The accuracy of proposed algorithms is represented by using ROC curve on Amazon dataset.

In [16], the authors proposed a random forest machine learning algorithm to analyze the sentiment reviews collected on Flipcart online shopping as negative, positive, or neutral. The results are expressed using SPYDER tool. The proposed model given better results compare to other machine learning approach support vector machine (SVM).

3. Proposed Work

Supervised machine learning mechanism is proposed to predict the sentiment of the target data. The supervised machine learning techniques could perform well in different domains with domain related, enough labelled data. The labelled data performs well when there exist all the domain related features that classify the sentiment.The mechanism of machine learning for sentiment text classification [17-18] was represented in figure 1. It includes mainly three elements, 1. Data analysis on input data, 2. Selection of classification model, 3. Solution evaluation on the chosen model. The data analysis was performed on the input labelled dataset is pre-handled to generate a statement in the form of features called feature extraction. The extracted features are weighted according to the selected feature representation algorithm. The weighted features are decreased by the feature collection for optimal data representation.

The feature extraction, weighting and selection of labelled data set is characterized appropriately for learning procedure. The representation of textual data of labelled data was represented as a vector of feature weight called vector space model. These features include unigrams, bigrams and n-grams [19]. Different classification models use different learning approaches. These learning methods are used to train a classification event based on the chosen classification model to predict or recognize the target text document class. The prediction of the chosen classifier produces the decision that define the probability or weights of the class of input vector. Classification model is selected based on the training function to recognize a target concept. Once a classification model is chosen, we can evaluate the model by measuring its performance. The model evaluation is associated to a document categorization objective. The different methods to measure their values are precision, recall and f-measure.

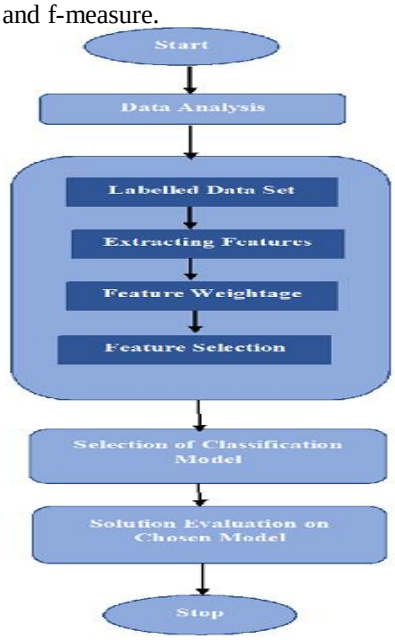

Fig. 1: Machine learning Approach for Sentiment text Classification

 https://doi.org/10.22362/ijcert/2021/v8/i12/v8i701

4. Approaches that effect the sentiment Analysis Performance

This section describes the supervised machine learning approaches that effect the sentiment analysis performance while dealing with real time data. Performance and accuracy of sentiment classification becomes a challenge while dealing with big data. The labelled training data of the model should contain all the domain related features that classify the sentiment. This will increase the prediction accuracy of real time unseen data. The performance accuracy of a sentiment categorization model is strongly dependent upon the type of words utilized in the corpus and type of characteristics designed for the algorithm. Extraction of variety of features plays an important role in classification algorithm. These features are domain related so the performance accuracy of the real time unseen testing data will affect if its domain does not belong to the training data domain. Because the probability of occurrences of features of training and testing data will be different.

Uneven distribution of training data for each class will also affect the performance accuracy. For example, if the training data was classified with three classes, the class weight should be equally distributed. One class with less data and another with more data will affect the performance accuracy. When dealing with real time data even distribution of class weight should be considered, because the training data with extremely high frequency of features will helps to classify the class. If there exist any common features belongs to the classes in the training data, in that case it chooses the probability of the feature of the class weight is more. To overcome this problem, we can balance the class weight such that that all the classes having reasonable number of records.

5. Result Discussion

Experiments are conducted on movie review dataset that consists of twenty-five thousand records with positive and negative polarity. Sentiment analysis is performed using Ensemble technique such as ensemble vote classifier with base classifiers as logistic regression, random forest and XGB classifiers. This ensemble vote classifier votes the class that is predicted by the maximum classifiers. The Sentiment analysis technique with ensemble vote classifier is performed with and without balanced class weight. The experimental results show the improved performance accuracy with balanced class weight compared to without balanced class weight. These techniques and approaches show drastic change in performance accuracy when dealing with real time data.

Table 1: Sentiment Analysis using Ensemble Technique without class weight on Training Data

Accuracy score: 0.868				
F1 score: 0.868				
precision		recall	f1-score	support
0 1	0.88 0.86	0.85 0.88	0.87 0.87	3112 3138

Table 2: Sentiment Analysis using Ensemble Technique with class weight on Training Data

Accuracy score: 0.8744				
F1 score: 0.8744				
precision		recall	f1-score	support
0 1	0.89 0.86	0.85 0.90	0.87 0.88	3112 3138

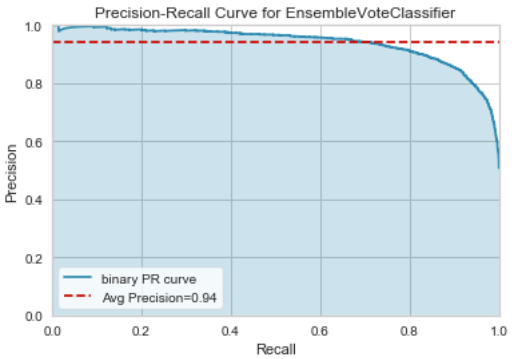

Fig 2: Precision and recall curve for Ensemble vote classifier without class weight

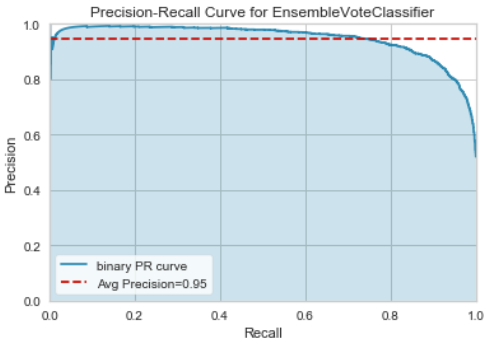

Fig 3: Precision and recall curve for Ensemble vote classifier with class weight

 https://doi.org/10.22362/ijcert/2021/v8/i12/v8i701

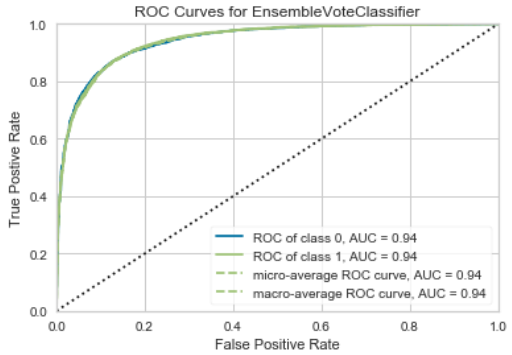

Fig 4: ROC curve for Ensemble vote classifier without class weight

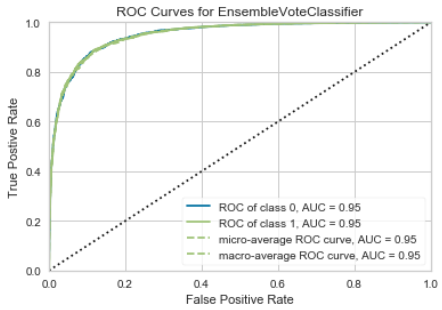

Fig. 5: ROC curve for Ensemble vote classifier with class weight

6. Conclusion

A novel ensemble vote classifies technique is proposed in this article for sentiment analysis using logistic regression, random forest and XGB classifiers. The proposed technique is applied on movie review data for sentiment analysis. The Sentiment analysis technique with ensemble vote classifier is performed with and without balanced class weight. The experimental results shown the improved performance accuracy with balanced class weight compared to without balanced class weight.

References

1. B. Pang and L. Lee, "Opinion mining and sentiment analysis", Foundation and Trends in Information Retrieval, vol. 2, 2008, pp. 1–135.
2. B. Liu , "Sentiment analysis and opinion mining", Morgan and Claypool Publishers, Synthesis Lectures on Human Language Technologies, vol. 5 2012, pp. 1–167.
3. Y. Mejova, "Sentiment Analysis: An Overview", 2009, retrieved from https://www.researchgate.net/publication/264840229_Sentiment_Analysis_An_Overview
4. Y. Zhang, and B. Wallace, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification", IJCNLP, 2017.
5. Nimirthi, Pradeepthi, P. Venkata Krishna, Mohammad S. Obaidat, and V. Saritha. "A framework for sentiment analysis based recommender system for agriculture using deep learning approach." In Social Network Forensics, Cyber Security, and Machine Learning, pp. 59-66. Springer, Singapore, 2019.
6. G. Preethi, P.V. Krishna, S. Obaidat Mohammad, V. Saritha and Y. Sumanth, " Application of Deep Learning to Sentiment Analysis for recommender system on cloud", *International Conference on Computer, Information and Telecommunication Systems (CITS)* , 2017, pp. 93-97.
7. Kavitha, Modepalli, and P. Venkata Krishna. "IoT-Cloud-Based Health Care System Framework to Detect Breast Abnormality." Emerging Research in Data Engineering Systems and Computer Communications. Springer, Singapore, 2020. 615-625.
8. Y. He and D. Zhou, "Self-training from labeled features for sentiment analysis", Information Processing and Management, vol. 47, 2011, pp. 606–616.
9. F. Xianghua, L. Guo, G. Yanyan and W. Zhiqiang, "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon", Knowledge Based Systems, vol.37, 2013, pp. 186–195.
10. Huang, C. F., Hsieh, T. N., Chang, B. R., & Chang, C. H. (2012, September). A comparative study of regression and evolution-based stock selection models for investor sentiment. In 2012 Third International Conference on Innovations in Bio-Inspired Computing and Applications (pp. 73-78). IEEE.
11. Hoque, M. T., Islam, A., Ahmed, E., Mamun, K. A., & Huda, M. N. (2019, February). Analyzing Performance of Different Machine Learning Approaches With Doc2vec for Classifying Sentiment of Bengali Natural Language. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.
12. Khan, S., Chopra, K., & Sharma, P. (2020, February). Brand Review Prediction using User Sentiments: Machine Learning Algorithm. In 2nd International Conference on Data, Engineering and Applications (IDEA) (pp. 1-8). IEEE.
13. Oyebode, Oladapo, FelwahAlqahtani, and Rita Orji. "Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews." IEEE Access 8 (2020): 111141-111158.
14. Aziz, A. A., Starkey, A., & Bannerman, M. C. (2017, September). Evaluating cross domain sentiment analysis using supervised machine learning techniques. In 2017 Intelligent Systems Conference (IntelliSys) (pp. 689-696). IEEE.
15. R. B. Shamantha, S. M. Shetty and P. Rai, "Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 21-25, doi: 10.1109/CCOMS.2019.8821650.

16. Karthika, P., R. Murugeswari, and R. Manoranjithem. "Sentiment Analysis of Social Media Network Using Random Forest Algorithm." 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). IEEE, 2019.
17. Mirończuk, Marcin &Protasiewicz, Jaroslaw. (2018). A Recent Overview of the State-of-the-Art Elements of Text Classification. Expert Systems with Applications. 106. 10.1016/j.eswa.2018.03.058.
18. Jonnavithula, S. K., Jha, A. K., Kavitha, M., & Srinivasulu, S. (2020, October). Role of machine learning algorithms over heart diseases prediction. In AIP Conference Proceedings (Vol. 2292, No. 1, p. 040013). AIP Publishing LLC.
19. Matthew Chang, Chung Keung Poon,Using phrases as features in email classification,Journal of Systems andSoftware,Volume 82, Issue 6,2009,Pages 1036-1045,ISSN 0164-1212,https://doi.org/10.1016/j.jss.2009.01.013.

