IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 11 Iss 11, 2022

Machine Learning-Based Sentiment Analysis Method in Data Mining with SVM Classifier

[1]Raja Ram Sah
Assistant Professor, Department of Computer Science and Engineering
Government Engineering College Jehanabad, Bihar
[2]Md Iftekhar Ahmad
Assistant Professor, Department of Computer Science and Engineering, Sitamarhi Institute of Technology, Sitamarhi, Bihar
[3]Manoj Kumar Sah
Assistant Professor, Department of Computer Science and Engineering, B. P. Mandal College of Engineering, Madhepura, Bihar

Abstract— Text orientation is classified as either positive or negative in sentiment analysis, which is viewed as a classification task. This study presents the findings of an experiment using benchmark datasets to train a sentiment classifier using Support Vector Machine (SVM). A new era that preserves the genuine nature of technology and digitalization has emerged as the globe has undergone a change. The market has changed at an astounding rate, thus it is imperative to take advantage of and inherit the benefits and prospects it offers. The emergence of web 2.0 has brought with it scalability and limitless reach, thus it would be disastrous for an organisation to ignore the new tactics in the competitive landscape this expanding virtual world has set along with its advantages. Organisations are now able to gather, classify, and analyse user evaluations and comments from microblogging sites about their services and products because to the advanced and sophisticated data mining techniques. The most classical traits were extracted using N-grams and various weighting schemes. Chi-Square weight characteristics are also investigated in order to choose informative features for the categorization. Chi-Square feature selection may significantly increase classification accuracy, according to experimental study.

Keywords— *Machine Learning, Sentiment Analysis Method, Data Mining, SVM Classifier, Chi-Square Feature Selection.*

INTRODUCTION

The process of determining writers' thoughts about particular entities is known as sentiment analysis, or opinion mining. The practice of looking through online product reviews to find the general consensus is known as sentiment analysis in reviews. Because sentiment analysis categorises a text's direction as either positive or negative, it is seen as a classification task. Apart from lexicon-based and linguistic methodologies, machine learning is another popular way for sentiment categorization. A wider range of study topics, such as customer service and product reviews, have used sentiment analysis. This paper presents studies with an open-source data mining software package for machine learning. According to research by, there is no one instrument or method that consistently produces the greatest results when conducting trials. But occasionally, some people perform better than others. The use of microblogging websites has increased significantly in recent years due to the quick development of mobile internet. Conversely, there is a growing trend of people sharing their opinions and experiences on goods and services. Furthermore, before deciding which new good or service to buy, people base their decisions on the opinions of prior clients. Similarly, businesses can use microblogging platforms (such as Facebook, Twitter, and others) to solicit feedback from customers about their goods and services. By investigating and evaluating the feedback, they can then make improvements to their targeted goods and services. It is not feasible to peruse

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 11 Iss 11, 2022

every review in a tweet, nonetheless. The studies were carried out utilising benchmark datasets by, selecting features using Chi-square and extracting features using a variety of term-weighting approaches. Support Vector Machines (SVM) has been employed in the process of classification.Precision, Recall, Accuracy, F Measure, and AUC were used to measure the outcomes and assess how effective the suggested strategy was. Many scholars have been focusing on creating automated methods and algorithms for text classification and sentiment analysis. The main goal of sentiment analysis is to categorise a given text into three categories: positive, negative, and neutral. There are now three fundamental methods for sentiment analysis available in the literature: hybrid (combining machine learning and lexicon), lexicon-driven, and machine learning-based. Various Lexicon driven sentiment analysis methods and techniques were examined by the authors in. Various machine learning methods that are applied to sentiment analysis have been thoroughly covered in. Additionally, to further improve the findings, researchers merged machine learning and lexicon-based approaches to create a hybrid framework that yielded even better results, as described in. Within the class of supervised machine learning algorithms is SVM. An algorithm for supervised machine learning must first be taught using pre-identified output classes (training data) before it can be used to categorise actual input data (test data). There are numerous annotated datasets covering various topics that are utilised by machine learning algorithms for sentiment analysis and categorization. The customer review dataset, the pros and cons dataset, the Amazon product review dataset, and the gender classification dataset are a few of these annotated datasets.

RELATED WORKS

Many pertinent studies have been developed in the field of sentiment analysis in recent years. The studies address a wide range of sentiment analysis-related topics, including corpus sizes and types, multilingual contexts, and dataset domains. The study carried out by focused on sentiment analysis of Twitter data. Regarding domains, has revisited the field of sentiment analysis using a collection of machine learning-based techniques for categorizing news articles about sports and cricket. Another domain-related issue was looked into by, who combined various pre-processing techniques to apply sentiment analysis to online movie reviews. Research by investigated a new field of study using Support Vector Machines (SVM) for testing various data sets from movie reviews, computer, hotel, or music-related topics, as well as opinions from digital cameras. Additionally, a lot of research has been done on sentiment analysis in multilingual contexts. For instance, Web forum posts in Arabic and English were subjected to sentiment categorization techniques. In addition to syntactic data, a broad range of stylistic traits in both Arabic and English were included in the trials. Enhancing accuracy and determining important characteristics for every sentiment class are the primary goals. To tackle these problems, an appropriate feature selection technique must be used to extract the valuable information prior to categorization. The classification performance can be improved if the characteristics are sturdy and dependable. In addition to lengthening computation times, an excessive number of characteristics reduce classification accuracy. Consequently, in order to increase accuracy and speed up computation in text classification tasks, feature selection is essential. One of the hottest academic issues of the day is the development and improvement of automated algorithms for sentiment extraction and analysis. A thorough examination and assessment of the most recent research on sentiment analysis using SVM is still necessary, despite the fact that numerous researchers have worked on sentiment analysis techniques using a variety of approaches (Lexical, Machine Learning, and Hybrid). It goes without saying that both the people who create or market these programmes and the new consumers who will purchase them depend on these

reviews. The authors outlined the suggested fixes for mining-related difficulties while also pointing out new and unresolved problems. A thorough survey of the literature is done in to assess the state of Arabic text mining at the moment. More than a hundred publications were chosen for this review from a variety of trustworthy sources, and they were then categorised based on their respective fields. A quantitative examination of a subset of articles is also carried out in terms of the type of publishing, the year, the category, and the contribution. A study of the literature on sentiment analysis and opinion mining of social issues was done by the researchers in. The following are a few relevant papers on sentiment analysis. The authors of carried out a thorough analysis of the literature on opinion mining from user evaluations in mobile app stores. The researchers emphasised the significance of mobile applications in the modern world and the rising need for user ratings of those apps.

RESEARCH PROTOCOL

The goal of this study is to collect useful data from the majority of pertinent sentiment analysis and opinion mining research articles that have been published in the last five years. According to a systematic literature review examines the gaps between several studies conducted within a specific time frame. The framework of a research protocol specifies the various procedures that must be followed in a specific order. In order to choose the most pertinent research articles with high quality metrics, this study uses a special approach that includes boundary lines and a specific structure. The steps in the research protocol/methodology for this study are as follows (Figure. 1):

- Determine the research questions.
- Choosing the query string's keywords,
- determining the search space, and
- stating the selection criteria
- Literature extraction using selection criteria
- Evaluating the quality of the literature that has been retrieved,
- extracting and synthesizing data, and presenting the results

Figure 1- Steps of SLR

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 11 Iss 11, 2022

SENTIMENT ANALYSIS APPROACH BY USING SUPPORT VECTOR MACHINE

(i) A Feature Based Approach - It created a method for Support Vector Machine (SVM)-based feature-based sentiment analysis. The dataset must go through five stages in the suggested model in order to reach the ultimate outcome. The classification of sentence levels is done first. Reviews that are neutral, negative, or have sentimental value should be the only ones kept. POS tagging will be used to filter out questions and comments that aren't reviews because retaining them would result in an unneeded expansion of the word dictionary and unintended scoring. The most crucial and difficult process is aspect extraction, which comes after sentiment categorization at the sentence level. Words with tags like NNS (noun plural), NN (noun), NNP (proper noun singular), etc. are extracted using POS tagging. The Stanford parser is used in the following step to extract the opinion words for aspects. SentiWordNet is then used to label the dataset. In order to determine the opinion regarding the entire product, the labelled dataset was subjected to an SVM classifier. SVM plots vectors in a three-dimensional virtual space and assigns the points in the testing data to certain categories, such as positive, negative, neutral, or any other predetermined groupings. The dataset used in this study came from user reviews of laptops made by a range of manufacturers, including Apple, Dell, Lenovo, HP, and others.

(ii) Target Based Polarity Phenomena- It described a subject-sensitive sentiment analysis method that takes into account tweet context in. The authors claim that using text purification techniques to input data prior to the classification process can enhance the outcome. Normalisation and vector representation of the input data are two aspects of text purification. It has been highlighted that subject-aware classification yields superior outcomes when compared to subject-unaware classification. If the uni-gram strategy is utilised in place of the bi-gram or n-gram approach, the results can be further enhanced. First, a Twitter dataset pertaining to the word "Obama" was chosen. Alchemy API, Tweet NLP, and NTLK were used to extract features from the tweets of the chosen dataset. Thirty percent of the dataset was used for training, while the remaining seventy percent was used for testing. To preserve the focus and context of the tweets, the gathered tweets were scanned for features. The features were then extracted and placed in a different dictionary called Keyword_Bundle along with their respective themes. This method contributed to the creation of the input matrix used by SVM to classify tweets more accurately. Then, "Apple" and "Movie Review" were chosen as two more datasets for comparison. The "Obama," "Movie Review," and "Apple" datasets yielded accuracy percentages of 85.00%, 84.00%, and 85.00%, respectively, for a cumulative accuracy of 85.60%.

(iii) Educational Data Catering Accreditation Process- It offered a method that divided the papers into several groups while taking into account a variety of factors. This study also took into account the current issues with document level sentiment analysis, including entity identification, subjectivity detection, and negation. The suggested paradigm was used to data mining in education. The student comments used as input for the evaluation of the faculty's performance. There were 5000 comments on the professors in the dataset. Questions, comments, and responses from social media platforms were removed from consideration as objective reviews with no bias. Two token groups were created from the reviews using the Java string tokenizer. Subsequently, special characters and certain pronouns that would have no practical significance in the classification itself. The collected data was represented numerically using TF-IDF, which is then utilised by the classifiers. Using the pre-processed dataset, two machine learning classifiers were used: Naïve Bayes and Support Vector Machine. For aspect-based document level sentiment analysis, the SVM and Naïve Bayes algorithms yielded 81.00% and 72.80% accuracy, respectively.

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 11 Iss 11, 2022

PROPOSED METHOD

This experiment aims to enhance SVM on benchmark datasets provided by Taboada Corpus and Pang Corpus. Preprocessing, feature extraction, feature selection, and classification phases make up the framework. In the subsection that follows, the success metric will also be briefly explained.

(i) Preprocessing Methods- The pre-processing tasks of tokenization, stop word removal, low-case conversion, and stemming will be applied to the datasets. The process of tokenizing a text involves dividing it into words, sentences, or other significant sections, or tokens. Stop words are terms like conjunctions, prepositions, etc. that are frequently used in texts but are not specific to any one subject. Lowercase conversion is an additional preprocessing step. Prior to the classification steps, all capital letters are typically changed to their lowercase equivalents. The final step in the stemming process is obtaining the stem and root of derived words. Porter Stem, which was first presented by, is the method of stemming English that is most frequently employed.

(ii) Datasets Description- For the preliminary experiments, we have used two label datasets which are 2000 positive and negative Movie Review Datasets from, and 400 positive and negative SFU Review Corpus Datasets from for the experiments.

(a) Pang Corpus- The corpus was prepared by to classify movie reviews collected from IMDb.com. The collection consists of 2000 reviews (1000 positive samples and 1000 negative samples).

(b) Taboada Corpus- This collection was prepared by that includes 400 opinions collected from the website Epinions.com divided into 200 reviews classified as ”recommended” (positive) and 200 as ”not recommended” (negative). The datasets contains reviews about product and services such as movies, books, cars, phones and etc.

(iii) Feature extraction- The process of turning an input data set into a set of features is called feature extraction. Selecting the appropriate features for extraction is essential since the machine learning process's features have a significant impact on its performance? The goal is to condense and convert the input data into a set of representation features (also known as features vectors) that the classifier can use effectively. However, one of our primary objectives is to apply several n-gram models that is, unigrams, bigrams, and trigrams in order to assess the impact of employing various n-gram systems.

(iv) Feature Selection- For feature selection, there are filter, wrapper, and embedding methods. Because of the classifier independence and the filters' comparatively short computation times, filter approaches were employed in the experiments.

(a) Chi-Square- Evaluation of filter attributes of In order to choose useful characteristics, Chi-Square weight features were employed, and a ranking mechanism was also used to exclude unnecessary features. CHI2 is one of the most often used feature selection techniques. The CHI2 test is a statistical tool used to assess the independence between two events. It is considered that the two events, X and Y, are independent if

$$p(XY) = p(X)p(Y). \quad (1)$$

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 11 Iss 11, 2022

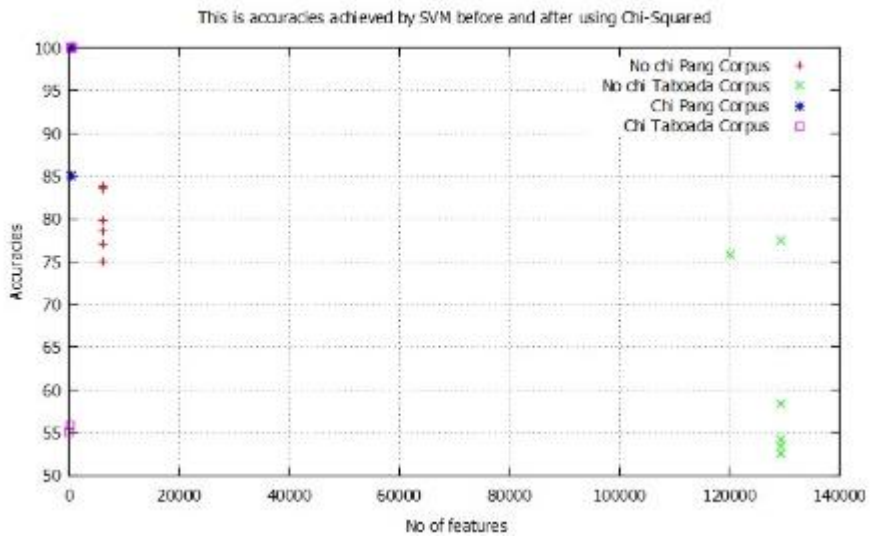

Figure 2- The correlation between accuracies and no of features before and after using Chi-Squared

It demonstrates how choosing the features according to their chi-squared statistical value assisted in lowering the text's dimensionality and noise level, enabling the classifier to perform well enough to be used for topic categorization. It also demonstrates that smaller feature sets and other variables that could impact classifier performance, like corpus size and domains, result in higher SVM prediction accuracies.

(v) Text Classification Method Selection- the Support Vector Machine (SVM) has been selected for the experiments' classification. Introduced by the support-vector machine is a learning machine for two-group classification tasks. It's employed to categorise the writings as either positive or negative. SVM's benefits, namely its capacity to handle big data, make it an effective tool for text classification. Another benefit of Support Vector Machines (SVM) is their ability to withstand sparse sets of examples and the fact that most problems can be solved linearly. Support Vector Machines have demonstrated encouraging outcomes in earlier sentiment analysis studies.

CONCLUSION

In the field of knowledge discovery, sentiment analysis is regarded as one of the hottest research subjects. Every day, a significant amount of internet data is added, encompassing anything from software and movie reviews to social media posts and comments. Through the application of sentiment analysis techniques, these data sources can yield valuable insights such as the ability to forecast election outcomes, gather user feedback regarding software, assess a brand's market reputation, and gauge public opinion prior to releasing a new product, among other things. For sentiment analysis, there are several methods available, including hybrid approaches that combine machine learning and lexicon-based techniques. SVM is one of the popular machine learning methods for polarity identification in text. Nowadays, researchers have presented numerous customised and integrated models for sentiment analysis and polarity identification in addition to traditional machine learning classification techniques. This paper focused on the SVM approach of sentiment analysis and offered a succinct and thorough evaluation of recent research. This study adhered to a methodical review process and, following a critical evaluation of a subset of publications, offered

responses to the research questions that were identified. It is advised to do a comparison analysis of the customised procedures using the same dataset in future work.

REFERENCES

[1] M. Ahmad, S. Aftab, S. S. Muhammad, and U. Waheed, "Tools and Techniques for Lexicon Driven Sentiment Analysis : A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 1, pp. 17–23, 2017.
[2] M. Ahmad, S. Aftab, and S. S. Muhammad, "Machine Learning Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, p. 27, 2017.
[3] M. Ahmad, S. Aftab, I. Ali, and N. Hameed, "Hybrid Tools and Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, 2017.
[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04, 2004, p. 168.
[5] X. Ding, X. Ding, B. Liu, B. Liu, P. S. Yu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," Proc. Int. Conf. Web search web data Min. - WSDM '08, p. 231, 2008.
[6] M. Ganapathibhotla and B. Liu, "Mining opinions in comparative sentences," Proc. 22nd Int. Conf. Comput. Linguist. - COLING '08, vol. 1, no. August, pp. 241–248, 2008.
[7] N. Jindal and B. Liu, "Opinion spam and analysis," Proc. Int. Conf. web search web data Min. 2008, pp. 219–230, 2008.
[8] A. Mukherjee and B. Liu, "Improving Gender Classification of Blog Authors," Proc. 2010 Conf. Empir. Methods Nat. Lang. Process., no. October, pp. 158–166, 2010.
[9] N. Genc-Nayebi and A. Abran, "A systematic literature review: Opinion mining studies from mobile app store user reviews," J. Syst. Softw., vol. 125, no. November, pp. 207–219, 2017.
[10] H. Al-Mahmoud and M. Al-Razgan, "Arabic Text Mining a Systematic Review of the Published Literature 2002-2014," 2015 Int. Conf. Cloud Comput., no. November, pp. 1–7, 2015.
[11] V. Singh and S. K. Dubey, "Opinion Mining and Analysis: A Literature Review," 2014 5Th Int. Conf. Conflu. Next Gener. Inf. Technol. Summit, pp. 232–239, 2014.
[12] A. A. Sheibani, "Opinion mining and opinion spam: A literature review focusing on product reviews," 2012 6th Int. Symp. Telecommun. IST 2012, pp. 1109–1113, 2012.
[13] N. T. Liu and J. Salinas, "Machine learning in burn care and research: A systematic review of the literature," Burns, vol. 41, no. 8, pp. 1636–1641, 2015.
[14] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," Know.-Based Syst., vol. 36, pp. 226–235, Dec. 2012.
[15] M. F. Porter, "Readings in information retrieval," K. Sparck Jones and P. Willett, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, ch. An Algorithm for Suffix Stripping, pp. 313–316.
[16] S. Gunal and R. Edizkan, "Subspace based feature selection for pattern recognition," Information Sciences, vol. 178, no. 19, pp. 3716–3726, 2008.
[17] T. L.Ladha, "Feature selection methods and algorithms," International Journal on Computer Science and Engineering (IJCSE), vol. 3, p. 5, 2011.
[18] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," Information Processing & Management, vol. 50, no. 1, pp. 104 – 112, 2014.

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 11 Iss 11, 2022

[19] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[20] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Machine Learning: ECML-98, ser. Lecture Notes in Computer Science, C. N´edellec and C. Rouveirol, Eds. Springer Berlin Heidelberg, 1998, vol. 1398, pp. 137–142.
[21] 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, ser. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86.
[22] E. Martinez Camara, V. Martin, M. Teresa, J. M. Perea Ortega, and L. A. Urena Lopez, "Tecnicas de clasificacion de opiniones aplicadas a un corpus en espanol," Procesamiento de Lenguaje Natural, vol. 47, pp. 163–170, 2011.
[23] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, Mar. 2002.