International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 10, October 2022)

Support-Vector Analysis for Sentiment Analysis

Nagma Parveen Ansari[1], Prof. Jayesh Jain[2]
Department of Computer Science & Engineering, Global Nature Care Sangathan's Group of Institutions, India

***Abstract*-- Sentiment analysis is a branch of natural language processing (NLP) that has a wide range of uses for identifying subjectivity and opinion in texts. In this essay, we will examine classifiers for sentiment analysis of user opinions toward political candidates through comments and tweets using Support Vector Machines (SVM), following the methodology used in the pioneering study by Pang, Lee, and Vaithyanathan. The objective is to create a classifier that analyses sentiment by classifying user comments as positive or negative. It allows us to divide text into interesting categories.**

***Keywords*-- Opining mining, Support vector machine, TF-IDF,Sentiment Classification.**

I. INTRODUCTION

Sentiment is essentially an idea or viewpoint that is based on feeling rather than logic. It is sometimes referred to as the expression of sensitive feeling in art and literature. It is a kind of subjective impression and not facts. Sentiment analysis, often known as opinion mining, is a task in natural language processing and information extraction that seeks to extract the writer's feelings as they are expressed in favourable or unfavourable remarks, queries, and requests by scrutinising a significant number of documents. Sentiment. The computational method of extracting, categorising, comprehending, and identifying the opinions conveyed in various items is known as sentiment analysis. It makes an effort to pinpoint the attitude people have toward a particular thing. It uses computational methods and natural language processing (NLP) to automatically extract or categorise sentiment from often unstructured content. In general, sentiment analysis seeks to ascertain a speaker's or writer's attitude toward a subject or the overall tone of a text.

The process of information being passed from one person to another, or word of mouth (WOM), is a significant factor in how customers choose which services or goods to purchase. WOM occurs when customers spread attitudes, views, goods, or services among others in commercial contexts. Social networking is the foundation for WOM communication. Sentiment analysis is currently driven by the dramatic rise in Internet usage and the interchange of public opinion in recent years. There is a vast amount of structured and unstructured data on the Web.

It is a difficult challenge to analyse this data in order to extract the public attitude and opinion. Sentiment analysis is useful for online blog posts, product reviews, endorsements, user opinions of political candidates, etc. The related sentiment analysis study conducted by various researchers is included in the next section. Then, using SVM, which has been demonstrated to be one of the most effective learning algorithms for text categorization [9], we suggest a method for sentiment analysis.

II. RELATED WORK

Many researchers are trying to combine the text mining and sentiment analysis as next generation discipline [3] [6]. In sentiment analysis document – level classification is most promising topic [9]. In Sentiment classification there are four different levels of sentiment analysis - sentence level, document level, phrase level, word level. Subjectivity and sentiment are both relevant properties of language. Subjectivity refers to linguistic expression of somebody's opinion, beliefs, speculations. Main task of subjectivity is to classify the contents in objective or subjective. Figure 1 shows the sentiment analysis and subjectivity analysis classification.

<table>
<tr><th>Sentiment Analysis</th><th>Subjectivity Analysis</th></tr>
<tr><td>Positive</td><td rowspan="2">Subjective</td></tr>
<tr><td>Negative</td></tr>
<tr><td>Neutral</td><td>Objective</td></tr>
</table>

Fig 1. Classification of sentiment analysis and subjectivity analysis

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 10, October 2022)

Theresa et al. [4] introduced phrase-level sentiment analysis, which decides the relative polarity of a statement based on whether the supplied expression is neutral or polar. For a large subset, the method automatically identifies the contextual polarity. Sentimental phrases that are truly superior to the baseline, but calculations take a long time. Yi et al[8] .'s proposed sentence-level polarity categorization, which attempts to categorise each sentence's positive and negative attitude. To capture numerous attitudes that could be present in a single sentence, phrase-level categorization can also be nested within sentence-level categorization. However, it doesn't matter how well the sentiment was predicted [1]. Consequently, the new strategy is introduced. In addition, Pang & Lee [3] developed a technique that classified sentences as subjective or objective before doing sentiment classification on the subjective part of the sentence. However, the outcomes showed that it is insufficient for forecasting entities' emotion. The most difficult and successful model for sentiment categorization that Turney [5] proposed is based on document level and uses two approaches: term counting and machine learning technique [1]. By calculating the positive and negative phrases, the term counting approach derives a sentiment metric. The sentiment classification problem is once again modelled as a statistical classification task in [3], where authors propose machine learning algorithms. Machine learning algorithms typically perform better than term-counting methods and have been tailored to more complex situations, including domain adaptation, multi-domain learning, and semi-supervised learning for sentiment analysis. Adjectival terms are seen by Whitelaw et al. [6] as a key indicator of the polarity of sentiment in textual reviews. This method is based mostly on identifying and evaluating the words or groups of terms that have received the highest ratings, such as "extremely good" or "very terrible," etc. The supervised learning techniques Wang et al [2] presented have been widely used and have successfully classified sentiments. It is heavily dependent on a vast volume of labelled data, which makes the process time-consuming and costly.

To address the issue with supervised learning methods, many semi-supervised learning approaches have been presented. Small amounts of labelled data must be combined with larger amounts of unlabeled data for semi-supervised algorithms to work. Support Vector Machine (SVM), a supervised learning approach that divides data into two categories by creating an N-dimensional hyper plane, was proposed by Vapnik [6]. G(x) is utilised by SVM [7] as the discriminating function,

$$g(x) = w^T f(x) + b \quad (1)$$

wherew is the weights vector, b is the bias , and f(x) denotes nonlinear mapping from input space to high-dimensional feature space. The parameters w and b are learned automatically on the training dataset following the principle of maximized margin by

$$\min_2 \frac{1}{2} W^T W + C \sum_{i=1}^{N} c_i \quad (2)$$

Where N denotes the slack variables and C denotes the penalty coefficient. Due to the dimension of feature space is quite large in text classification task, the classification problem is always linearly separable [1,4] and therefore linear kernel is commonly used.

III. Proposed Work

An overview of sequential steps and techniques commonly used in sentiment classification approaches, as shown in Figure 1.Parts of speech is a model which aims to classify roles that means according to parts of speech has also been explored. In this model , information is used as part of a feature set which leads to sentiment classification on a dataset.

The model parts of speech is supposed to be the significant indicator of sentiment expression and which works on subjectivity detection that represents the close relationship between presence of adjectives and sentence subjectivity. But, many experimental results show that using only adjectives as features leads to worse performance.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 10, October 2022)

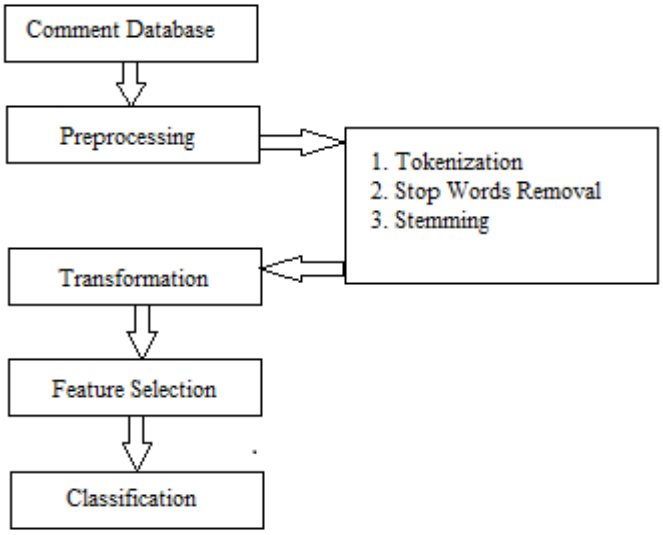

Fig. 2 Steps and techniques used in sentiment classification.

1. Text preprocessing, first: Preparing and cleaning a batch of data in preparation for classification is known as pre processing. The following is the premise of having the data appropriately pre-processed: lowering the text's noise level should help the classifier perform better and classify data more quickly, enabling real-time sentiment analysis.

2. Tokenization: Tokenization is the process of breaking up a character sequence into smaller units called tokens while also deleting certain characters like punctuation marks. A token is a specific instance of a character sequence that is compiled into a practical semantic processing unit.

3. Eliminating stop words: A stop-list is the conventional name for a group or list of stop words. Typically, it is language-specific, though it could include terms. One stop-list for each language may be present in a search engine or other natural language processing system, or there may be one stop-list that is multilingual. The following are some of the more commonly used stop words in English: "a," "of," "the," "I," "it," "you," and "and." These words are classified as 'functional words,' meaningless words. The message can be communicated more effectively when analysing the contents of natural language by omitting the functional words. Therefore, it is useful to eliminate words that appear excessively frequently but provide no information for the task.

4. The process of stemming involves returning derivative words to their stem, or root, form. Stemming algorithms or stemmers are typical names for stemming programmes. Simple and quick, this method involves using a stemmer to search up the inflected form in a lookup database. The drawback is that the table must clearly list every inflected form. For instance, "developed," "development," and "developing" are all shortened to "develop."

B. Transformation

The weight of each word in the corpus is calculated with the help of TF-IDF, so that it is easy to determine what words in the corpus of documents might be more favorable to use in a further processing. TF-IDF calculates [9] values for each word in a document defined as below –

$$w_d = f_{w,d} * \log(|D| f_{w,D})$$

D is collection of documents ,w represents words, d is individual document belongs to D,|D| is size of corpus, $f_{w,d}$is number of times w appears in d,$f_{w,D}$is number of documents in which w occurs in D.

C. Feature Selection

Feature Selection is used to make classifiers more efficient by reducing the amount of data to be analyzed as well as identifying relevant features to be considered in classification process. Ideally, feature selection stage will refine features, which are input into a classification / learning process.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 10, October 2022)

- Identify the parts of corpus to contribute to positive and negative sentiment.
- Join these parts of corpus in such a way that the document falls into one of these polar categories.

D. Classification

Goal of text classification is to classify data into predefined classes. Here they are positive and negative classes. Text classification is supervised learning problem.

First step in text classification is transforming document which is in string format into format suitable for learning algorithm and classification task. In information retrieval it is found that word stem works well as representation unit. This leads to attributed value representation of text. Each word corresponds to feature with, number of times word occurs in document, as its value. Words are considered as features only if they are not stop words (like "and", "or", etc). Scaling the dimension of feature with IDF improves the performance[12].

SVM- Support vector machines are universal learners[12]. Remarkable property of SVM is that their ability to learn can be independent of dimensionality of feature space. SVM measures the complexity of Hypothesis based on margin that separates the plane and not number of features[12].

SVM learning Algorithms for Text Categorization -

SVM has defined input and output format. Input is a vector space and output is 0 or 1 (positive/negative).

Text document in original form are not suitable for learning. They are transformed into format which matches into input of machine learning algorithm input. For this preprocessing on text documents is carried out.Then we carryout transformation.Each word will correspond to one dimension and identical words to same dimension. As mentioned before we will see TF-IDF for this purpose. Now a machine learning algorithm is used for learning how to classify documents, i.e. creating a model for input-output mappings. SVM has been proved one of the powerful learning algorithm for text categorization[12].

SVM Benefits]-

1. *High Dimension Input Space* - while text classification we have to deal with many features (may be more then 1000). Since SVM uses over fitting protection[12], which does not depend on number of features so they have ability tohandle large number of features.
2. *Document Vector Space*- despite the high dimensionality of the representation, each of the document vectors contain only a few non-zero element[12]. More Text Categorization problems are linearly separable[12].

SVM Characteristics-

1. ML algorithms typically use a vector-space (attribute-value) [10] representation of examples, mostly the attributes correspond to words. However word-pairs or the position of a word in the text may have considerable information, and practically infinitely many features can be constructed which can enhance classification accuracy.
2. Categories are binary, but generally documents are not assigned so precisely. Often a document D is said to belong a little to category X1 and a bit to category X2, but it does not fit well into any of the two. It probably would require a new category, as it is not similar to any of the documents seen before.
3. Number of words increase if we increase the number of documents. Heap's law[11] describes how the number of distinct words increase if number of document increases.
4. Representations use words as they are in texts. However, words may have different meanings, and different words may have the same meaning. The proper meaning of a word can be determined by its context i.e. each word influences the meaning of its context. However, the usual (computationally practical) representation neglect the order of the words. Task of SVM is to learn and generalize the input-output mapping. In case of text categorization input is set of documents and output is their respective class. Consider spam filter as example input is an email and output is 0 or 1(either spam or no spam)[10].

SVM Evaluation-

Text categorization systems may make mistakes. To compare different text classifiers for deciding which one is better, performance measures are used. Some of these measures the performance on one binary category, others aggregate per-category measures, to give an overall performance. TP, FP, TN, FN are the number of true/false positives/ negatives[13]. The most important per-category measures for binary categories are [13]

- Precision: p=TP/(TP+FP)
- Recall: r= TP/(TP+FN)

The most important averages are: micro-average[13], which counts each document equally important, and macro-average, which counts each category equally important.

IV. Conclusion And Future Works

With the use of TF-IDF and SVM, we examined information retrieval and preprocessing strategies in this study. Additionally, we investigate Support Vector Machine, which can be utilised to determine the polarity of textual commentary, for text categorization. According to the study, SVM recognises some textual traits such as a) high dimensional feature space, b) few irrelevant features, and c) sparse instance vector. A performance evaluation of SVM utilising recall and precision is also given in the paper. Various findings demonstrate that SVM performs well on text categorization when compared to ANN. SVM reduces the requirement for feature selection because of its capacity to generalise high dimensional feature space.

REFERENCES

[1] Ms. K. Nirmala Devi, Ms. K. Mouthami,Dr. V. Murali Bhaskaran' Sentiment Analysis and Classification Based on Textual Reviews',2012.

[2] Li, S., Wang, Z., Zhou, G., & Lee, S.Y.M. , 'Semi-supervised learning for imbalanced sentiment classification', In Proceedings of international joint conference on artificial intelligence, pp. 1826–1831,2012.

[3] Pang, B., & Lee, L., 'A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts', In Proceedings of the association for computational linguistics, pp. 271–278, 2004.

[4] Theresa Wilson, JanyceWiebe, Paul Hoffmann, 'Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis', In Proceedings Of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP),Vancouver, pp. 347–354 2004.

[5] Turney,'Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews', In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417-424, 2002.

[6] Vapnik, V., 'The Nature of Statistical Learning Theory', Springer-Verlag, pp. 863-884, 2000.

[7] Yang. Y, X. Liu, A re-examination of text categorization methods, in: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), ACM, New York, NY, USA, pp. 42–49,1999.

[8] Yi, J., Nasukawa, T., Niblack,W., &Bunescu, R. , 'Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques', In Proceedings of the 3rd IEEE international conference on data mining (ICDM 2003), USA, pp. 427– 434,2003.

[9] Rodrigo Moraes, Joao Francisco Valiati,Wilson P. GaviaoNeto,'Document-level sentiment classification : An empirical comparison between SVM and ANN',Expert Systems with Applications 40 621-633,2013.

[10] HumaLodhi, Craig Saunders, John Shawe-Taylor, NelloCristianini, Chris Watkins: Text Classification using String Kernels, The Journal of MachineLearning Research, Volume 2, pp. 419-444,2002.

[11] Heaps' law: http://en.wikipedia.org/wiki/Heaps'_law

[12] Thorsten Joachims: Text categorization with support vector machines: learningwith many relevant features, Proc. of ECML-98, 10th European Conference on Machine Learning, Springer Verlag, Heidelberg, DE,pp. 137-142,1998.

[13] FabrizioSebastiani: Machine learning in automated text categorization, ACM Computing Surveys (CSUR), Vol. 34 Issue 1, ACM Press, New York, NY, USA, pp. 1-47,2002.