Sentiment Analysis and Classification Based On Textual Reviews

Ms.K.Mouthami
PG Student,Dept of CSE
Kongu Engineering College
Perundurai, Erode, Tamil Nadu
mouthamik@gmail.com

Ms.K.Nirmala Devi
Assistant Professor,Dept of CSE
Kongu Engineering College
Perundurai, Erode, Tamil Nadu
k_nirmal@kongu.ac.in

Dr.V.Murali Bhaskaran
Principal,
Dhirajlal Gandhi College of Technology
Omalur,Salem,Tamil Nadu

Abstract- **Mining is used to help people to extract valuable information from large amount of data. Sentiment analysis focuses on the analysis and understanding of the emotions from the text patterns. It identifies the opinion or attitude that a person has towards a topic or an object and it seeks to identify the viewpoint underlying a text span. Sentiment analysis is useful in social media monitoring to automatically characterize the overall feeling or mood of consumers as reflected in social media toward a specific brand or company and determine whether they are viewed positively or negatively on the web. This new form of analysis has been widely adopted in customer relation management especially in the context of complaint management. For automating the task of classifying a single topic textual review, document-level sentiment classification is used for expressing a positive or negative sentiment. So analyzing sentiment using Multi-theme document is very difficult and the accuracy in the classification is less. The document level classification approximately classifies the sentiment using Bag of words in Support Vector Machine (SVM) algorithm. In proposed work, a new algorithm called Sentiment Fuzzy Classification algorithm with parts of speech tags is used to improve the classification accuracy on the benchmark dataset of Movies reviews dataset.**

Index Terms- **Sentiment analysis, opinion mining, Text classification, Support Vector Machine, Term weighting, Sentiment Fuzzy Classification, parts of speech tags.**

1. INTRODUCTION

Text Mining is used to extract previously unknown information from different written resources. A key element is used to link together the extracted information to form new facts or new hypotheses to be explored further by more conventional means of experimentation.

A. Sentiment Analysis

Sentiment analysis is the process used to determine the attitude/opinion/emotion expressed by a person about a particular topic. Sentiment analysis or opinion mining uses natural language processing and text analytics to identify and extract subjective information in source materials. The rise of social media such as blogs and social networks has fuelled interest in sentiment analysis. In order to identify the new opportunities and to manage the reputations, business people usually view the reviews/ ratings/ recommendations and other forms of online opinion. This allows to not only find the words that are indicative of sentiment, but also to find the relationships between words so that both words that modify the sentiment and what the sentiment is about can be accurately identified. Scaling system is used to determine sentiment for the words having a positive, negative and neutral sentiment. It also analyzes the subsequent concepts to understand the words and how they relate to the concept.

B. Learning Methods

The different learning types are:
Supervised learning**:** Learning classifier from training data and assign class labels to test data.
Unsupervised learning: Learning without training data.
Semi-supervised learning: Amalgamate both labeled and unlabeled training data. The Sentiment learning uses Machine Learning or Lexicon based learning.

C. Classification of textual review

Classification is a supervised procedure that learns to classify new instances based on learning from a training set of instances that have been properly labeled with the correct classes. An algorithm that implements classification, especially in a concrete execution is classifier. The piece of input data is formally termed an instance, and the categories are termed classes. Text Classification (TC) is one of the prime techniques to deal with the textual data. TC systems are used in a number of applications such as, filtering email messages, classifying customer reviews for large e-commerce sites, web page classification for an internet directory, evaluating exams paper answers and organizing document databases in semantic categories. Rodrigo et al (2012) for automating the task of classifying a single topic textual review, document-level sentiment classification are used for expressing a positive or negative sentiment. The propose system uses fuzzy set theory because it provides a more straightforward way to represent the interior fuzziness in sentiment. The rest of this paper organized as follows. In section 2 describe the related works are reviewed. Section 3 describes the proposed algorithm. Section 4 deals with the experimental setup and Section 5 conclude the discussion.

II. RELATED WORK

The Various research groups are exploring the ways to use Text mining and sentiment analysis as next generations paradigm shift. Document level classification is most promising topic in Sentiment analysis.

A. SENTIMENT CLASSIFICATION

Sentiment analysis can be performed at four different levels word level, phrase level, sentence level, and document level[3]. Wiebe et al [6] proposed Subjectivity and meaning are both important properties of language. Word hypotheses are

- Subjectivity is a property; it can be associated with word senses.
- Subjectivity annotations are directly [4] used to word sense disambiguation.

Each document is modeled as a sequence of observations (words) and underlying states, leads to increase in time. Theresa et al [13] proposed Phrase-level sentiment analysis. It determines whether an expression is neutral or polar and then disambiguated the polarity of the polar expressions. In this approach, automatically identify the contextual polarity for a large subset of achieving results, sentiment expressions that are significantly better than baseline but it takes more time. Pang et al [8] proposed a seminal approach in sentiment classification. Essentially, they conclude that machine learning techniques, like NB and SVM, do not achieve accuracy as good on sentiment classification as on traditional topic-based categorization. The classification accuracy resulting uses only unigrams as features. Yi et al [19] proposed Sentence level polarity categorization attempts to classify positive and negative sentiments for each or whether a sentence is subjective or objective. There has also been work on phrase level categorization in order to capture multiple sentiments that may be present within a single sentence. In this approach, we cannot accurately predict the sentiment, to overcome this problem we go for the next approach. Pang and Lee [9] also proposed classifying sentences as being either subjective or objective, and then apply sentiment classification on the subjective portion of the text. It is not sufficient for identifying sentiment of entities. Turney [14] proposed document-level sentiment classification there are two kinds of approaches: term-counting approaches and machine learning approaches. Term-counting approaches usually involve deriving a sentiment measure by calculating the total number of negative and positive terms. Pang and Lee [9] proposed machine learning approaches recast the sentiment classification problem as a statistical classification task. Compared to term-counting approaches, machine learning approaches usually achieve better performance, and have been adapted to more complicated scenarios, such as domain adaptation, multi-domain learning and semi-supervised learning for sentiment classification. Whitelaw et al [15] proposed considering adjectival expressions as an important indication of the sentiment polarity in textual reviews. Sentiment classification based on extracting and analyzing appraisal groups such as “very good” or “not terribly dreadfully”. Wang et al [5] proposed supervised learning methods have been widely employed and proven effective in sentiment classification. They normally depend on a large amount of labeled data, which usually involves high cost in labour and time. To overcome this problem, various semi-supervised learning methods are proposed to effectively utilize a small scale of labeled data along with a larger amount of unlabeled data. Semi-supervised methods [17] for sentiment classification are to utilize prior lexical knowledge in conjunction with the labeled and unlabeled data and to employ some bootstrap techniques, such as self-training and co-training. Vapnik [15] proposed SVM are a group of supervised learning methods that performs classification by constructing an N-dimensional hyper plane that optimally separates the data into two categories. Suzuki et al [12] proposed SVM model using a sigmoid kernel function is equivalent to a two-layer, perception neural network. SVM[7] has been shown to perform very good on a wide variety of classification problems that require large scale input space, such as handwritten character recognition, face detection, and most importantly in this case, text categorization. Rodrigo et al [11] proposed for automating the task of classifying a single topic textual review, document-level sentiment classification is used for expressing a positive or negative sentiment. Supervised learning methods consist of two stages, extraction/selection of informative features and classification of reviews by using learning models like SVM. Nirmala Devi [6] et al proposed predicting the hotspots based on sentiment analysis in online forums. Accuracy in the classification of positive or negative movie review is less in sentiment analysis. Based on the people opinion, the idea about the movie may vary that is the same movie may be good or bad depend upon the opinion of the people.SVM [18] uses g(x) as the discriminate function,

$$g(x) = w^T f(x) + b \tag{1}$$

where w 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

$$\min \frac{1}{2} W^T W + C \sum_{i=1}^{N} c_i \tag{2}$$

where N denotes the slack variables and C denotes the penalty coefficient. The above problem directly, it is converted to an equivalent quadratic optimization problem by Lagrange multipliers. The training sample $(\overline{x_i}, y_i)$is called a support vector. Due to the dimension of feature space is quite large in text classification tasks, the classification problem is always linearly separable [13] and therefore linear kernel is commonly used.

III. PROPOSED WORK

An overview of steps and techniques commonly used in sentiment classification approaches, as shown in Figure 1. Part of speech model in which a document is represented as a vector, whose entries correspond to individual terms of a vocabulary. Part-of-speech information is supposed to be a significant indicator of sentiment expression. The work on subjectivity detection [3] reveals a high correlation between the presence of adjectives and sentence subjectivity.

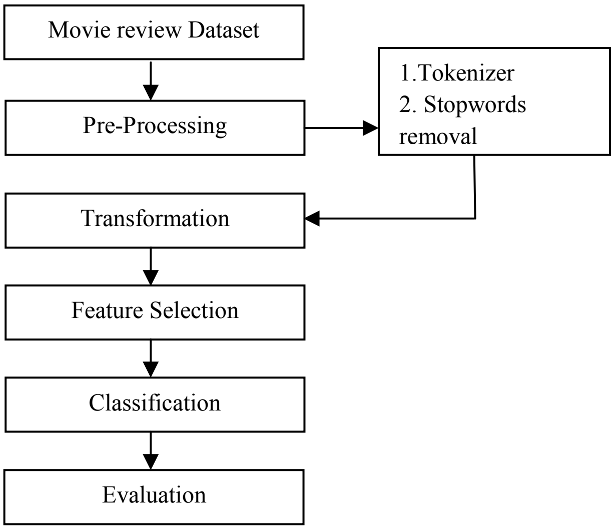

Fig 1. steps and techniques used in sentiment classification

Indeed, in the study [4], the experimental results show that using only adjectives as features actually results in much worse performance than using the same number of most frequent words.

A. Text Preprocessing

Text pre-processing techniques are divided into two subcategories.

1. Tokenization: Textual data comprises block of characters called tokens. The documents are separated as tokens and used for further processing.

2. Removal of Stop Words: A stop-list is the name commonly given to a set or list of stop words. It is typically language specific, although it may contain words. A search engine or other natural language processing system may contain a variety of stop-lists, one per language, or it may contain a single stop-list that is multilingual. Some of the more frequently used stop words for English include "a", "of", "the", "I", "it", "you", and ”and” these are generally regarded as 'functional words' which do not carry meaning. When assessing the contents of natural language, the meaning can be conveyed more clearly by ignoring the functional words. Hence it is practical to remove those words which appear too often that support no information for the task. If the stop word removal is applied, all the stop words in the particular text file will not be loaded. If the stop word removal is not applied, the stop word removal algorithm will be disabled when the dataset is loaded.

B. Text Transformation

The score of each sentence in the source document is calculated by sum of weight of each term in the corresponding sentences. The weight of each term is calculated by multiplication of TF and IDF of that word based on adjective word extracted from Parts of speech tags. The TF and IDF are defined as

TF (t) = Number of times the adjective term occurs in document(d)/Total Number of adjective in document(d)(3)

$$IDF(t)=\log\{ND/DF(t)\} \quad (4)$$

Here ND is total number of document in the document collection and DF (t) is number of documents in which adjective term (t) occurs in the document collection.

C. FEATURE SELECTION

Many statistical feature selection methods for document level classification can also be used for sentiment analysis. The simplest statistical approach for feature selection is to use the most frequently occurring words in the corpus as polarity indicators. The majority of the approaches for sentiment analysis involve a two-step process:

- Identify the parts of the document to contribute the positive or negative sentiments.
- Join these parts of the document in ways that increase the odds of the document falling into one of these two polar categories.

D. SENTIMENT FUZZY CLASSIFICATION

Sentiment polarity is vague with regard to its conceptual extension. There is not a clear boundary between the concepts of “positive”, “neutral” and “negative”. To better handle such intrinsic fuzziness in sentiment polarity, we apply the fuzzy set theory to sentiment classification. To do so, we first redefine sentiment classes as three fuzzy sets, and then apply existing fuzzy distributions to construct membership functions for the three sentiment fuzzy sets. A fuzzy set is defined by a membership function. These functions can be any arbitrary shape but are typically triangular or trapezoidal. In our formulation, the entire opinionated documents under discussion are represented as a sorted set, denoted by *X*, in terms of their opinion weight (calculated by TF-IDF).

X = [Min (Opinion weight(Si)), …, Max(Opinion weight(Si))]

Where, i={1,…,n}, Min(Opinion weight(Si)) and Max(Opinion weight(Si)) denotes the respective minimum and maximum opinion weight.

1. Positive sentiment fuzzy set: If X is a collection of sentiment opinions denoted by *x*, then a positive sentiment fuzzy set $\overline{P}$ in X can be defined as a set of ordered pairs are

$$\overline{P} = \{(x, \mu_{\overline{P}}(x)) \mid x \in X\} \quad (5)$$

where $\mu_{\overline{P}}(x)$denotes the membership function of *x* in $\overline{P}$ that maps *X* to the membership space *M*.

We select the rise semi-trapezoid distribution as the membership function of the positive sentiment fuzzy set, namely

$$\mu_{\overline{P}}(x) = \begin{cases} 0, & x<c \\ \frac{x-c}{d-c}, & c \le x \le d \\ 1, & x>d \end{cases} \quad (6)$$

where x denotes the opinion weight of a document under discussion. The adjustable parameters c and d can be defined as

$$c = \text{Min}(x_i) + \lambda_1 (\text{Max}(x_i) - \text{Min}(x_i)/k) \quad (7)$$

$$d = \text{Min}(x_i) + \lambda_2 (\text{Max}(x_i) - \text{Min}(x_i)/k) \quad (8)$$

Max (x_i) and Min (x_i) denote the respective minimum and maximum values within X. λ_1, λ_2 and k are parameters. Here we set λ_1= 5.2, λ_2 = 5.4, and k = 10.

*2. Neutral sentiment fuzzy set: I*f X is a collection of sentiment opinions (denoted by *x*), then a neutral sentiment fuzzy set $\overline{E}$ in X can be defined as a set of ordered pairs are

$$\overline{E} = \{(x, \mu_{\overline{E}}(x)) \mid x \in X\} \quad (9)$$

where $\mu_{\overline{E}}(x)$ denotes the membership function of *x* in $\overline{E}$ that maps X to the membership space M.

As shown in Formula (8), select the semi-trapezoid distribution as the membership function of the neutral sentiment fuzzy set.

$$\mu_{\overline{E}}(x) = \begin{cases} 0, & x<c \\ \frac{x-c}{x-d}, & c \le x \le d \\ 1, & d \le x \le e \\ \frac{f-x}{f-e}, & e \le x < f \\ 0, & x \ge f \end{cases} \quad (10)$$

where x denotes the opinion weight of a document under test. c, d, e and f are adjustable parameters that can be defined as

$$e = \text{Min}(x_i) + m_2 (\text{Max}(x_i) - \text{Min}(x_i)/k) \quad (11)$$

$$f = \text{Min}(x_i) + \lambda_2 (\text{Max}(x_i) - \text{Min}(x_i)/k) \quad (12)$$

Max(x_i) and Min(x_i) denotes the respective minimum and maximum values within X. λ_1, λ_2,m_1, m_2 and k are parameters, Here we set λ_1 = 5.2,λ_2 = 5.5, m_1 = 5.26, m_2 = 5.33, and k = 10.

*3. Negative sentiment fuzzy set: I*f X is a collection of sentiment opinions (denoted by *x*), then a negative sentiment fuzzy set $\overline{N}$ in X can be defined as a set of ordered pairs are

$$\overline{N} = \{(x, \mu_{\overline{N}}(x)) \mid x \in X\} \quad (13)$$

where $\mu_{\overline{N}}(x)$ denotes the membership function of *x* in $\overline{N}$ that maps X to membership space M. To represent the membership function of the negative sentiment fuzzy set, we employ the drop semi-trapezoid distribution namely

$$\mu_{\overline{N}}(x) = \begin{cases} 1, & x<c \\ \frac{d-x}{d-c}, & c \le x \le d \\ 0, & x>d \end{cases} \quad (14)$$

where x denotes the opinion weight of a document under discussion. The adjustable parameters c and d can be defined as Max (x_i) and Min (x_i) refer to the corresponding minimum and maximum values in X. λ_1, λ_2, and k are parameters. Here we set λ_1=5.2, λ2=5.3 and k=10.

E. Parameters for evaluation

In the context of classification, True Positives (TP), True Negatives (TN), False Negatives (FN) and False Positives (FP) are used to compare the class labels assigned to documents by a classifier with the classes the items actually belongs to. True positive means, which are truly classified as the positive terms. True positives (TP) are examples that the classifier correctly labeled as belonging to the positive class. False positive (FP) are examples which were not labeled by the classifier as belonging to the positive class but should have been. True Negative (TN) is examples that the classifier correctly labeled as belonging to the negative class. True Negative means, which are truly classified as the Negative terms. At last there is False Negative (FN), which is an example which was not labeled by the classifier as belonging to the negative class but should have been. Other evaluation measures like precision, recall, F-measure, specificity and accuracy can easily be calculated from these four variables.

Table 1.Contegency table

		Correct labels	
		Positive	Negative
Classified labels	Positive	TP(True positive)	FP(False positive)
	Negative	FN(False negative)	TN(True negative)

1. Accuracy: A common measure for classification performance is accuracy, or its complement error rate. Accuracy is the proportion of correctly classified examples to the total number of examples, while error rate uses incorrectly classified instead of correctly. However, one should be careful to use only accuracy when one is using skewed data. This is because when one class occurs significantly more than the other, the classifier might get higher accuracy by just labelling all examples as the dominant class then what it gets when it tries to classify some with the other class.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

2. Precision and recall: Precision and recall are two widely used metrics for evaluating performance in text mining, and in other text analysis field like information retrieval. They can be seen as extended versions of accuracy, and by using a combination of these measures the problem with skewed data for classifiers dissipates. Precision is used to measure exactness, whereas recall is a measure of completeness. Precision is the number of examples correctly labeled as positive divided on the total number that are classified as positive, while recall is the number of examples correctly labeled as positive divided on the total number of examples that truly are positive. This is shown in the following formulas.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

3. *F-Measure:* F-Measure is the harmonic mean of precision and recall. This gives a score that is a balance between precision and recall. F-Measure combines them into one score for easier usage. This is important because it might be better to optimize the system to favors either the precision or the recall if one of these has a more positive influence on the final result of the trading simulation than the other.

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{18}$$

IV.EXPERIMENTAL SETUP

A. Datasets

The widely used document level sentiment classification dataset is Cornell movie-review corpora[1], introduced in [9], However, in this paper focus only on positive and negative reviews. No single author could have more than 20 reviews in the original data set.

B. Database of opinions

LETHAL WEAPON is a light-weight action thriller that quickly degenerates to an extremely stupid movie. While a detective whose name I don't remember is investigating the murder of a friend's daughter, he and hi new partner, Riggs, a suicidal burned-out ex-Vietnam sniper, stumble onto a heroin ring.
This movie is so loaded with cliches, ex-CIA men turned to crime, a tough looking guy holding his arm in flame to prove his loyalty to Fearless Leader, kidnapped daughters, and so on, that there is no room left fo any genuine suspense. What is left is an adventure movie with no guts or heart.
If you look hard for something worthwhile in this movie, there is a vague chemistry between Glover an Gibson that provides a couple of good laughs. But that chemistry is erratic at best, and doesn't come close t saving the film.

Fig 2.Database of opinions

ID	Positive	Negative
1	good	bad
2	great	worst
3	efficient	inefficient
4	fair	unfair
5	supportive	not
6	recommend	flop
7	suggest	low
8	acceptable	below
9	average	vague
10	above	sleepy
11	nice	bored

Fig 3.Database of sentiment

1.http://www.cs.cornell.edu/people/pabo/movie-review-data/.

V.CONCLUSION

In sentiment analysis, it is difficult for human to predict the movie review. To resolve this, the document-level sentiment classification is used in the existing system. It determines whether an opinion document (movie review) is positive or negative or neural sentiment. It can be approximately classifies the sentiment using the Bag of words. To make the classification accurate, parts of speech can be used. A new algorithm called Sentiment Fuzzy Classification algorithm is proposed to improve classification accuracy on the benchmark dataset of Movies reviews.

REFERENCES

1. Benamara.F, C. Cesarano, A. Picariello, D. Reforgiato, V.S. Subrahmanian, Sentiment analysis: adjectives and adverbs are better than adjectives alone,in: Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2007,vol.8,pp 234-245.
2. Christopher D. Manning, Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, USA, 1999,vol.3,pp 12-34.
3. Hatzivassiloglou . V, J. Wiebe, Effects of adjective orientation and gradability on sentence subjectivity, in: Proceedings of the International Conferenceon Computational Linguistics (COLING), 2000, pp. 299–305.
4. Janyce Wiebe and Rada Mihalcea (2006), 'Word Sense and Subjectivity', Joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics, pp. 73-99.
5. Li, S., Wang, Z., Zhou, G., & Lee, S.Y.M. (2011), 'Semi-supervised learning for imbalanced sentiment classification', In Proceedings of international joint conference on artificial intelligence, pp. 1826–1831.
6. Nirmala Devi K., & Murali Bhaskaran V (2012),"Online forums hotspot prediction base on Sentiment Analysis" in Journal of Computer Science, vol. 8, No.8, pp 1219-1224.
7. Nirmala Devi K., & Murali Bhaskaran V (2012),"Text sentiments for forums hotspots detection" In Proceedings of International Journal of Information Science and Techniques, vol. 2, Issue 3, pp 53-62.
8. Pang, B., Lee, L., & Vaithyanathan, S. (2002), 'Thumbs up ? sentiment classification using machine learning techniques', In Proceeding of the conference on empirical methods in natural language processing, pp. 79–86.
9. Pang, B., & Lee, L. (2004), 'A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts', In Proceedings of the association for computational linguistics, pp. 271–278.
10. Riloff.E, J. Wiebe, T. Wilson, Learning subjective nouns using extraction pattern bootstrapping, in: Proceedings of the Conference on Natural LanguageLearning (CoNLL), 2003, pp. 25–32.
11. Rodrigo Moraes, Joao Francisco Valiati , Wilson P. Gaviao Neto (2012), 'Document-level sentiment classification: An empirical comparison between SVM and ANN', journal of Expert Systems with Applications, UNISINOS, pp. 1-13.
12. Suzuki, Y., & Fukumoto, F. (2001), 'Learning Lexical Representation for Text Categorization', Proceedings of 2nd NAACL, pp. 23-67.
13. Theresa Wilson, Janyce Wiebe, Paul Hoffmann.(2005), '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.
14. Turney (2002), '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.
15. Vapnik, V. (2000), 'The Nature of Statistical Learning Theory', Springer-Verlag, pp. 863-884.
16. Wang, Y., & Wang, X. (2005), 'New Approach to Feature selection in Text Classification', Proceedings of the 4th International Conference on Machine Learning and Cybernetics. IEEE, pp. 145-189.
17. Whitelaw, C., Garg, N., & Argamon, S. (2005), 'Using appraisal groups for sentiment analysis', In Proceedings of the 14th ACM international conference on information and knowledge management, pp. 625–631.
18. 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, 1999, pp. 42–49
19. Yi, J., Nasukawa, T., Niblack,W., & Bunescu, R. (2003), '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.