Classification of Sentiment Reviews using POS based Machine Learning Approach

K. S. Kalaivani, R. Felicia Grace, M. Aarthi, M. Boobeash
Department of Computer Science and Engineering,
Perundurai, Erode,
Tamil Nadu

***Abstract*:- Most of the companies which are active in social media and online marketing sites use the public forum to promote their brands and services. Manual classification of reviews into positive or negative polarity is a time - consuming process. In order to improve the efficiency, the classification process is automated. The reviews which are usually unstructured are pre – processed and POS tagged for classification of human sentiments. The words that are mostly sentiment bearing are extracted and weighted. Supervised Machine Learning algorithms such as Naïve Bayes(NB), Support Vector Machine(SVM) and Maximum Entropy(ME) are employed to classify reviews. Further, performance of the classification algorithms are examined based on the performance parameters such as Accuracy, Precision, Recall and f-measure.**

1.INTRODUCTION

Sentiment analysis otherwise known as opinion mining is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions and their associated attributes [1]. In the current scenario it is extremely useful because it provides information about any product from different reviews, blogs, and comments. Opinions that are mined from social networking and online marketing sites can be valuable in deriving meaningful information which act as an important source for further analysis and improved decision making. There are different levels of sentiment classification like Document level, Sentence level and Entity level classifications. The task at the document level is to classify whether a whole opinion document expresses a positive or negative sentiment. Sentence level determines whether each sentence expresses a positive, negative, or neutral opinion. Aspect level focuses on all expressions of sentiments present within given document and the aspect to which it refers [2]. There are many ways to implement Sentiment Analysis. Ultimately, it is a text classification problem and can be broken down into two main areas. Supervised learning technique involves the construction of a "Classifier" and the problem has been studied intensively. The Classifier is responsible for categorizing texts into either a positive, negative or neutral polarity [3]. Unlike supervised learning, unsupervised learning process do not need any label data; hence they cannot be processed at ease. The reviews are usually unstructured and it may contain some unwanted information, so the reviews are subjected to pre - processing which removes symbols, punctuations numbers etc., POS tagging is done on the pre - processed reviews to extract the features. The tagged reviews are converted into a matrix of numbers which is given as an input to classifiers. The performance of the reviews is evaluated using performance metrics. The main contribution of the paper is as follows:

i. Different machine learning algorithms are proposed for the classification of Health data set using POS based machine learning techniques viz., Adjective and Adverb, Noun, Verb and combination of Adjective, Adverb, Noun and Verb.
ii. For converting text data into a matrix of numbers different weighting techniques like Binary weighting(BIN), Term Frequency(TF) and Term Frequency – Inverse Document Frequency(TF-IDF) weighting are done.
iii. Three supervised machine learning techniques such as Naïve Bayes(NB), Support Vector Machine(SVM) and Maximum Entropy(ME) are used for classification purpose.
iv. The performance of the classification algorithms is analysed using metrics such as Accuracy, Precision, Recall and f-measure.

The results obtained in this paper indicate, the higher values of accuracy when compared with studies made by other authors of the previous work.
The structure of the paper is defined as follows: Section 2 presents literature review. In Section 3, the proposed approach is explained. Section 4, concludes the paper and presents the scope for future work and also the performance evaluation of the proposed approach is carried out.

2.LITERATURE REVIEW

Pang et.al (2002) have considered the aspect of sentiment classification based on categorization study, with positive and negative sentiments Pang, Lee, and Vaithyanathan (2002). They have under- taken the experiment with three different machine learning algorithms, such as, NB, SVM, and ME. The classification process is undertaken using the n-gram technique like unigram, bigram, and combination of both

unigram and bigram. They have used bag-of- word features framework to implement the machine learning algorithms. As per their analysis, NB algorithm shows poor result among the three algorithms and SVM algorithm yields the result in a more convincing manner [4][5].

Liu et.al and Chen (2015) have proposed different multi-label classification on sentiment classification. They have used eleven multilevel classification methods compared on two micro-blog dataset and also eight different evaluation matrices for analysis. Apart from that, they have also used three different sentiment dictionary for multi-level classification. According to the authors, the multi-label classification process perform the task mainly in two phases i.e., problem transformation and algorithm adaptation [6].

Basant Agarwal et.al., (2013) incorporated the information of POS-based sentiment-rich phrases in a machine-learning algorithm that determines the semantic orientation of a given text. Bi-tagged phrases were used as features in combination with unigram features for sentiment classification. Joint feature vectors of unigrams and bi-tagged phrases have high dimensions consisting of noisy and irrelevant features. A feature selection method was used to select only relevant features from the feature vector. Experimental results show that the combination of prominent unigrams and bi-tagged phrases outperforms other features for sentiment classification in a movie review dataset [7].

Zhi-Hong Deng, Kun-Hu Luo, Hong-Liang Yu (2013) used Term weighting strategy to assign weights to terms to performance of sentiment analysis and other text mining tasks. A supervised term weighting scheme based on two basic factors namely, importance of a term in a document (ITD) and importance of a term for expressing sentiment (ITS), to improve the performance of analysis were proposed to improve the performance. Seven statistical functions were employed to learn the ITS of each term form maintaining documents with category label. The experimental results show that the proposed method outperformed the existing methods and produced the best accuracy [8].

Salvetti *et.al.* , (2004) have discussed on Overall Opinion Polarity (OvOp) concept using machine learning algorithms such as NB and Markov model for classification Salvetti, Lewis, and Reichenbach. In this paper, the hypernym provided by wordnet and Part Of Speech (POS) tag acts as lexical filter for classification. Their experiment shows that the result obtained by wordnet filter is less accurate in comparison with that of POS filter. In the field of OvOp, accuracy is given more importance in comparison with that of recall. In their paper, the authors presented a system where they rank reviews based on function of probability. According to them, their approach shows better result in case of web data [4].

Beineke *et.al.* , (2002) have used NB model for sentiment classification. They have extracted pair of derived features which are linearly combinable to predict the sentiment Beineke, Hastie and Vaithyanathan. In order to improve the accuracy result, they have added additional derived features to the model and used labelled data to estimate relative influence. They have followed the approach of Turney which effectively generates a new corpus of label document from the existing document Turney. This idea allows the system to act as a probability model which is linear in logistics scale. The authors have chosen five positive and negative words as anchor words which produce 25 possible pairs and they used them for the coefficient estimation [9].

Peter D. Turney (2002) proposed a simple unsupervised learning algorithm for classifying reviews as recommended or not recommended. The classification of a review was predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. The semantic orientation of a phrase was calculated as the mutual information between the given phrase and the word “excellent” minus the mutual information between the given phrase and the word “poor” [10].

3. PROPOSED WORK

The workflow of the proposed work is given in Figure.1 The first step involves the collection of reviews. Amazon Multi Domain Review dataset is used. The reviews are pre - processed to remove the symbols, punctuations, numerals, etc. POS is done to extract adjective, adverb, nouns and verbs which are further used as features for Matrix construction. Different weighting techniques like Bin, TF, TF-IDF are used and the resultant matrix is given as input to classifiers like Naïve Bayes, Support Vector Machine and Maximum Entropy.

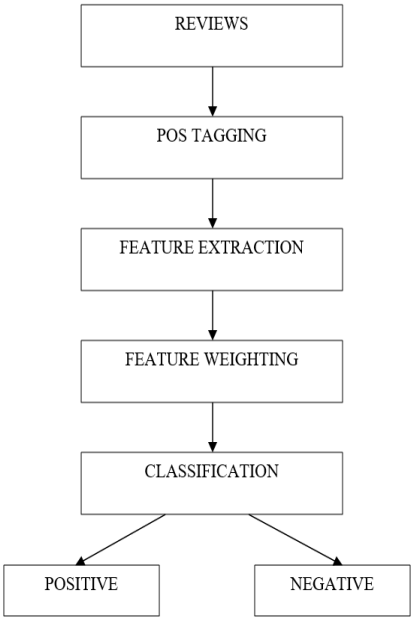

Figure.1 Flow diagram of the proposed work

3.1 REVIEWS

The dataset contains product reviews metadata from Amazon spanning from May 1996- July 2014 containing different domains like camera, mobile, kitchen, electronics, DVD, Books, fashion, Health, cosmetics etc. This dataset includes product reviews from customers. Health domain is used. The dataset is collected from the website www.cs.jhu.edu. The sample dataset used throughout the explanation contains 2000 samples of which 1800 are training data and 200 are test data.

3.2 POS TAGGING

In corpus linguistics, part-of-speech tagging also called grammatical tagging or word-category disambiguation is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context. By gaining the POS tags of each and every word in the corpus, the words that add the sentiments larger can be chosen and considered for analysis. POS weighting scheme works in the principle that more important parts of speeches like verb, adjective, adverb etc. are given higher weights when compared to other POS tags [11]. Some of the POS tags that convey sentiments are,

Part of Speech Tag	Part of Speech Category
NN NNS NNP NNPS	Noun, singular or mass Noun, plural Proper noun, singular Proper noun, plural
JJ JJR JJS	Adjective Adjective, comparative Adjective, superlative
VB VBD VBG VBN VBP VBZ	Verb, base form Verb, past tense Verb, gerund or past participle Verb, past participle Verb, non-3rd person singular present Verb, 3rd person singular present
RB RBR RBS	Adverb Adverb, comparative Adverb, superlative

Table.1 List of Tag that convey sentiments

Consider the example given in figure,

I feel happy

Figure.2 Sample Review before POS tagging

I/NN feel/VB happy/JJ

Figure.3 Sample Review after POS tagging

3.3 FEATURE EXTRACTION

Feature Extraction creates a set of features by decomposing the original data. A feature is a combination of attributes that is of special interest and captures important characteristics of the data. The feature becomes a new attribute. Feature extraction can also be used to enhance the speed and effectiveness of supervised learning [12]. After POS tagging, not all the tags convey sentiments. Feature sets such adjective and adverb, noun, verb and combination of adjective, adverb, noun and verb are extracted which are mostly sentiment bearing. For example in Figure.3 the words “feel” and “happy” which are tagged as “VB” and “JJ” conveys the sentence as a positive polarity. Therefore such features are extracted.

3.4 FEATURE WEIGHTING

Feature weighting or term weighting is the assignment of numerical values to terms that represent their importance in a document in order to improve effectiveness. It considers the relative importance of individual words in a sentiment analysis system, which can improve system effectiveness, since not all the terms in a given document collection are of equal importance. Weighting the terms is the means that enables the system to determine the importance of a given term in a certain document or corpus. It suggests the importance of the word to the document and whole corpus. Different types of weighting schemes that are used in this sentiment analysis system are,

- Binary Weighting
- Term Frequency Weighting
- Term Frequency- Inverse Document Frequency Weighting

i. Binary Weighting (BIN)

Binary weighting is the process of representing the occurrences of terms in the document with the help of either 0 or 1. When the term is present in the document, it is given a term weight of 1 and when the term is absent in the document, it is given a term weight of 0. In this way, a document term matrix is built using binary term weighting scheme. Consider 100 documents containing the word cat 5 times.

Without considering the count of how many times the word has appeared in the document, we just denote the presence of the word as 1 and its absence as 0 in binary term weighting scheme.

ii. Term Frequency Weighting (TF)

Term Frequency weighting scheme takes the number of occurrences of a particular term in a document. In this scheme, the total number of occurrences of a particular term in the document is counted and the occurrence count is used as the term weights. Term Frequency measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Consider a document containing 100 words wherein the word cat appears 3 times. The term frequency for cat is then (3/100) = 0.03.

iii. Term Frequency – Inverse Document Frequency Weighting (TF-IDF)

The TF-IDF weight is composed of two terms: the first computes the normalized Term Frequency which is the number of times a word appears in a document, divided by the total number of words in that document and the second is the Inverse Document Frequency, computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears. Inverse Document Frequency diminishes the weight of terms that occur very frequently in the document and increases the weight of terms that occur rarely. Certain terms like “is”, “of”, and “that”, may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones. Consider a document containing 100 words wherein the word cat appears 3 times. The frequency (i.e.,TF) for cat is then (3/100) = 0.03. Now, assume we have 10 million documents and the word cat appears in one thousand of these. Then, the inverse document frequency (i.e.,IDF) is calculated as log(10,000,000/1000) = 4. Thus, the TF-IDF is the product of these quantities: 0.03 * 4 = 0.12.
In this paper, the above three weighting schemes are used to convert the text format into matrix of numbers which is then considered as an input to supervised machine learning algorithms

3.5 CLASSIFICATION

Supervised learning is the machine learning task of inferring a function from labelled training data. In machine learning and statistics, classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. The different types of classification that are used in the work are,

- Naïve Bayes Method
- Support Vector Machine Method
- Maximum Entropy Method

i. Naïve Bayes Method

Naive Bayes method is used for both classification as well as training purposes. It works on Bayes theorem of probability to predict the class of unknown data set. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’. Naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters. Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods.

ii. Support Vector Machine Method

Support Vector Machine is a supervised machine learning algorithm which can be used for both classification or regression tasks. However, it is mostly used in classification problems. In this algorithm, each data item is plotted as a point in n- dimensional space with the value of each term being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiate the two classes very well. Support Vectors. Are the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes with the help of a hyper plane or a line. It works really well with clear margin of separation. It is effective in high dimensional spaces. It is effective in cases where number of dimensions is greater than the number of samples. It uses a subset of training points in the decision function called as support vectors, so it is also memory efficient.

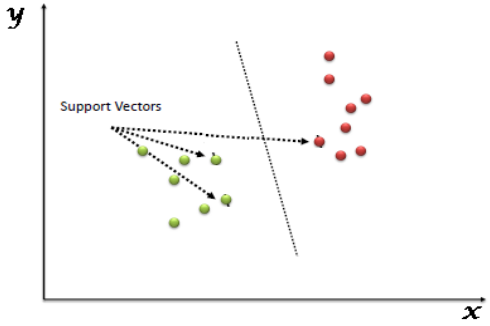

Figure.4 Support Vector Machine

The above diagram for support vector machine shows the hyper plane that splits the dataset into two domains.

iii. Maximum Entropy Method

The Maximum Entropy classifier is closely related to a Naïve Bayes classifier, except that, rather than allowing each feature to have its independency, the model uses search-based optimization to find weights for the terms that maximize the likelihood of the training data. The features defined for a Naïve Bayes classifier are easily ported to a MaxEnt setting, but the MaxEnt model can also handle mixtures of Boolean, integer and real-valued features. Entropy is the quantity that measures the uncertainty of a distribution. Among all the models that fit the training data, the one with the maximum entropy is selected.

4. RESULTS AND CONCLUSION

The performance of the classification algorithms can be evaluated using a matrix called confusion or contingency matrix. From classification point of view, terms such as "True Positive(TP)", "False Positive (FP)", "True Negative(TN)", "False Negative (FP)" are used to compare label of classes in this matrix. True Positive represents the number of reviews which are positive and also classified as positive by the classifier, where False Positive indicates the number of reviews which were incorrectly classified as positive. Similarly, True Negative represents the reviews which are negative also classified as negative by the classifier, where False Negative are the number of reviews which are incorrectly classified as negative. Based on the values obtained from confusion matrix, other parameters such as "precision", "recall", "f-measure", and "accuracy" are found out for evaluating performance of any classifier.

	Correct Labels	
	Positive	Negative
Positive	TP(True Positive)	FP(False Positive)
Negative	FN(False Negative)	TN(True Negative)

Table.2 Confusion Matrix

Precision: It measures the exactness of the classifier result. It is the ratio of number of examples correctly labelled as positive to total number of positively classified example. It is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

$$\text{Precision} = TP/(TP+FP)$$

Recall: It measures the completeness of the classifier result. It is the ratio of total number of positively labelled example to total examples which are truly positive. It is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

$$\text{Recall} = TP/(TP+FN)$$

F-Measure: It is the harmonic mean of precision and recall.

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

Accuracy: It can be calculated as the ratio of correctly classified example to total number of examples.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

This work attempts to improve the classification accuracy by properly selecting the feature sets. Then the classification algorithms are applied to it. It is observed that for adjective + adverb and combination of adjective, adverb, noun and verb improves the accuracy remarkably better but when noun classification alone is carried out the value of accuracy decreases.

The proposed work also has some limitations,

i. Most of the reviews may contain symbols like ☺, 👍,👎,☹ which are good in sentiment bearing but in POS tagging it is not taken into consideration as it is being removed during pre-processing.

ii. If a sentence contains "not good" the overall tone is negative polarity, but in feature extraction we extract the word "good"(JJ) and classify it as positive polarity. In such cases the classified result will be incorrect.

iii. In order to give stress on a word, it is observed that some per- sons often repeat the last character of the word a number of times such as "greatttt, Fineee". These words do not have a proper meaning; but they may be considered and further processed to identify sentiment.

All of above mentioned limitations may be considered for the future work, in order to improve the quality of sentiment classification.

REFERENCES

1. https://medium.com/retailmenot- engineering/sentiment-analysis-series-1-15- min-reading-b807db860917
2. https://www.cs.uic.edu/~liub/FBS/Sentiment Analysis-and-OpinionMining.pdf
3. https://www.growthaccelerationpartners.com/blog/sentiment-analysis/

4. http://daneshyari.com/article/preview/38196 7.pdf
5. https://www.sciencedirect.com/science/articl e/pii/S095741741630118X
6. https://www.researchgate.net/profile/Abinash_Tripathy/publi cation/299420336_Classific ation_of_Sentiment_Reviews_using_N-gram_Machine_Learning_Approach/links/5 9f05ac00f7e9beabfc6744a/Classification-of- Sentiment-Reviews-using-N-gram-Machine- Learning-Approach.pdf
7. https://vdocuments.site/ieee-2013-ieee-13th- international-conference-on-data-mining-workshops-icdmw 589b8469008c5.html
8. https://www.hindawi.com/journals/mpe/201 1/872347/
9. https://www.researchgate.net/publication/299420336_Classifi cation_of_Sentiment_Revi ews_using_N-gram_Machine_Learning_Approach
10. http://nparc.nrc- cnrc.gc.ca/nparc/eng/view/object/?id=4bb7a 0c8-9d9b-4ded-bcf6-fdf64ee28ccc
11. http://acopost.sourceforge.net/
12. https://docs.oracle.com/cd/B14117_01/data mine.101/b10698/4descrip.htm