Use of NLP Based Combined Features for Sentiment Classification

K.S.Kalaivani, S.Kuppuswami, C.S.Kanimozhiselvi

Abstract*: Sentiment analysis is the technique of automatic detection of the belief or the mood of an author towards a certain subject in textual form. To extract the opinion present in text, the machine needs expertise in the area of natural language processing. In this paper, machine learning based document-level sentiment classification is performed on Amazon product reviews to classify them as positive and negative. Two NLP based feature extraction techniques (Word Relation and POS based) are used in this study to determine the features that are sentiment bearing. The features are extracted as basic features (unigrams, bigrams and trigrams) and their combinations (unigrams+bigrams, unigrams+trigrams, unigrams+bigrams+trigrams). In order to identify the features that are most informative and to bring down the computational time of the classification algorithms, feature selection techniques are used. Performance of independent and combined feature sets is assessed using accuracy, precision, recall and F-measure. From the experiments conducted, it is observed that combined features outperformed independent features using Boolean Multinomial Naive Bayes (BMNB) classifier.

Index Terms*: Document-level Sentiment Classification, Information Gain, NLP based combined features, Weighted Frequency and Odds.

I. INTRODUCTION

With the recent development of the web, many people prefer to purchase products online and post their experience for different products they purchase. It has become a common practice for the manufacturers to analyze the customer's opinions about their products to improve user satisfaction. The consumers are also interested to find out whether others like or dislike a product before purchasing. Due to the availability of extremely large number of reviews for a given product, it is becoming harder and harder for the people to understand and evaluate the opinions manually. So, there is a need to automate this process.

Sentiment analysis also called as Opinion Mining is a branch of study that takes the natural language text as input and aims to extract the sentiments present in the text using some computational methods [9]. The two most commonly used approaches for sentiment analysis are semantic orientation and machine learning approaches [3,4]. The problems faced by semantic orientation approaches is that i) Non-availability of a single corpus which can provide polarity of features based on the context and domain and ii) there is no available knowledge base containing affective knowledge (instead they contain only general information which is not sufficient for detecting the polarity) [1]. Machine learning approaches perform better than semantic orientation approaches.

Sentiment Classification using machine learning approaches involves three main tasks: Feature Extraction, Feature Selection and Classification. The reviews available in unstructured format have to be converted into a suitable format for further processing. Feature Extraction which is the task of identifying sentiment bearing words from the text, results in a huge dimension of features. So, feature selection is used to extract only the significant features removing the irrelevant and noisy features. The reduced feature set is then given as input to the machine learning algorithms to improve the accuracy of classification task.

The remainder of this paper is prepared as follows: Section 2 discusses about the other studies related to this work. The proposed work on document level sentiment classification using word relation and POS based features are presented in Section 3. Section 4 focuses on the dataset used, evaluation metrics and experimental results. Finally, Section 5 provides some conclusions of the present work.

II. RELATED WORK

N-gram and tag-based features have been widely used in the earlier works for sentiment classification using machine learning approaches [6,8,14]. N-gram features are words where 'N' indicates how many words exist in a feature. Tag-based features have their words tagged with their Part-of-Speech (POS) or Sentiwordnet scores. The tags can either be used as part of features or just for selecting the features.

Pang, Lee and Vaithyanathan introduced machine learning based sentiment analysis and used unigrams, bigrams and adjectives as features [13]. The authors used SVM, Maximum Entropy and Naive Bayes for classification on movie review dataset. The conclusion was that binary weighting gave higher accuracy than term frequency when used with unigrams and SVM gave the best accuracy. Mejova and Srinivasan used different types of Part-of-Speech (POS) tagged features like adjectives, adverbs and nouns and analyzed their performance for supervised sentiment analysis. Their experimental results show that adjectives out performs other POS-tagged features when used individually.

Pak and Paroubek used subgraphs extracted from the dependency tree of a parsed sentence for constructing the feature vector [12]. They performed experimentations on movie reviews and decided that the subgraph-based features along with SVM classifier gave the best performance. Nguyen et al. improved the accuracy of document-level sentiment analysis by merging new rating-based features along with

Revised Manuscript Received October 05, 2019

K.S.Kalaivani, Department of CSE, Kongu Engineering College, Perundurai, India.

Prof.S.Kuppuswami, Principal, Kongu Engineering College, Perundurai, India.

unigrams, bigrams, and trigrams on the movie review dataset [10]. Hung and Alfred used word-based features like unigrams, POS based features, sentiwordnet features and phrase-based features like bigrams and trigrams [8]. The authors have concluded that phrase-based features outperform word-based features as the latter disregard the sequence and distort the meaning of the original text.

Machine learning based sentiment classification face the problem of dealing with huge number of features. In order to improve the performance of sentiment analysis with respect to accuracy and execution time, feature selection methods are used to eliminate the irrelevant and redundant features. Many researchers have worked on feature selection methods like Document Frequency, Mutual Information, Information Gain, to reduce the size of the feature vector [1,5,11,15]. Wang et al. introduced a novel feature selection technique based on boolean and frequency information called Fisher's discriminant ratio [16]. The experiments show that frequency-based Fisher's discriminant ratio outperforms IG using SVM classifier. Agarwal and Mittal proposed the hybridized IG and rough set-based feature selection method for sentiment analysis [1]. The authors have also used Categorical Probability Proportional Difference (CPPD) and minimum Redundancy and Maximum Relevancy (mRMR) feature selection techniques to improve the accuracy of sentiment analysis. In most of the earlier work, IG outperformed other feature selection metrics.

Feature weighting schemes play a vital role in sentiment classification as they assign weights to the features based on their sentiment importance in order to improve the classification results. Earlier works have used term weighting schemes like Binary, Term Frequency and Term Frequency - Inverse Document Frequency [13]. All the three weighting schemes are used for analysis in this work. Earlier works have adopted Naive Bayes (NB), Support Vector Machine (SVM), Maximum Entropy (ME) and neural network classifiers for sentiment analysis [7,15]. But, the performance of SVM classifier greatly exceeds other classifiers for sentiment classification. However, Agarwal and Mittal [1] proved that Boolean Multinomial Naive Bayes (BMNB) classifier outperforms SVM classifier for sentiment analysis. So, in this paper BMNB and SVM classifiers are used to classify the different features extracted using lexical and linguistic approaches.

Most of the earlier works used both N gram and POS based features independently. This work uses their combinations to further enhance the performance of machine learning classifiers. The combined features results in huge dimension of features and hence the feature selection methods like IG and WFO are used to select the most optimal ones which are used for efficient model construction.

III. PROPOSED WORK

Classification of sentiment reviews using machine learning approaches involves the following tasks: preprocessing, feature extraction, feature selection and classification. The reviews are in textual (unstructured) format and hence they need to be converted to a format compatible for further processing.

A. Preprocessing

The reviews are first tokenized followed by stop word removal, special character removal and case normalization. Stemming is then performed to reduce the words to their root form. When a negation word ("no", "isn't", "not", "didn't", etc.) is present in a sentence, it changes the sentiment of the sentence. So, the tag 'NOT' is added to every word after the negation word to the first punctuation mark. For example, in the sentence, "I do not like this product", the polarity of the word "like" is reversed by "not". So, after handling negation the sentence is written as "I do NOT_like, NOT_this, NOT_product".

B. Feature Extraction

Different feature sets are extracted in order to explore the influence of each feature type for sentiment classification.

Word relation-based features:

Unigrams are simple bag-of-word (BoW) features extracted by removing the noisy characters and spaces between any two words. Bigrams and trigrams are type of N-gram features in which a feature is made up of two and three consecutive words in the sentence respectively. Unigrams ignore the word structure and sequence, and hence distort the meaning of the sentence. But bigrams and trigrams are ways of preserving the sentence structure up to some extent. In this approach, unigrams, bigrams and trigrams are used as basic features and their combinations (unigrams+bigrams, unigrams+trigrams and unigrams+bigrams+trigrams) are used as combined features. The description of word relation-based feature sets used in this work is given in Table I.

POS based features:

Part of Speech (POS) denotes the grammatical category of a word and this information can be used to extract the sentiment-rich features. For extracting these features, the reviews (need not be preprocessed) are first tagged using the Stanford POS tagger. It has been investigated in earlier works that adjectives and adverbs have more subjectivity. But verbs and nouns also hold sentiment information and they are useful for further analysis. Hence, the words belonging to the categories namely adverbs, adjectives, nouns and verbs are extracted as unigrams. Bigrams and trigrams are constructed as *-adjective/adverb/noun/verb, adjective/adverb/verb-* respectively, where * denotes any POS category. The description of the POS based feature sets used in this work is given in table I.

Feature Selection

Feature selection is done to decrease the number of features in the feature vector. It also increases the classifier's performance in terms of accuracy and computational speed.

Information Gain (IG):

IG measures the bits of information gained for class prediction of an arbitrary text document by estimating the presence or absence of a feature in that text document. It measures the change in entropy value when a feature is present vs absent [17]. For a given feature fi, IG is calculated as follows.

$$IG(f_i) = -\frac{A_i+B_i}{N}\log\frac{A_i+B_i}{N} + \frac{A_i}{N}log\frac{A_i}{A_i+C_i} + \frac{B_i}{N}log\frac{B_i}{B_i+D_i} \quad (1)$$

where Ai is the number of reviews that contain the feature fi and also b

Bi is the number of reviews that do not contain the feature fi but belong to class Ci

Ci is the number of reviews that contain the feature fi but do not belong to class Ci

Di is the number of reviews that neither contain the feature fi nor belong to class Ci

N is the total number of reviews in the training collection i.e., N= Ai+Bi+Ci+Di

Ni is the number of reviews that belong to class Ci

Weighted Frequency and Odds (WFO):

Features that possess high document frequency (Ai or Ci) and high category ratio ((Ai)/(Bi) or (Ci)/(Di)) are said to be good features. However, using any one single measure does not perform well in selecting the set of optimal features. A feature selection method called Weighted Frequency and Odds tunes the importance accordingly and is estimated as follows [17].

$$WFO(f_i) = \left(\frac{A_i}{N_i}\right)^{\lambda}\left(log\ \frac{A_i\,(N - N_i)}{C_i N_i}\right)^{1-\lambda} \quad (2)$$

where λ is the parameter used to tune the weight between document frequency and category ratio and its value varies from 0 to 1. When the value of λ is 0, the formula becomes equal to Mutual Information (category ratio) and when the value of λ is 1, the formula becomes equal to document frequency. The value of λ is varied from 0 to 1 in steps of 0.1 during each run of 10-fold cross validation to get the best performance.

C. Feature Weighting

The relevant features selected are then given weights using Binary, Term Frequency (TF), Term Frequency – Inverse Document Frequency (TF-IDF) weighting schemes. In binary weighting, a feature is given a weight '1' if it is present in a document and '0' otherwise. Term Frequency measures the frequency of occurrence of a particular feature in a document. TF-IDF is computed as the product of TF and IDF where

$$IDF = \frac{\text{Total number of documents}}{\text{Number of documents in which the feature occurs}} \quad (3)$$

Classification

Support Vector Machine (SVM) classifier has been widely used for sentiment analysis and proved to provide excellent results. Agarwal and Mittal included less redundant and more informative features in the feature vector and showed that the BMNB classifier combined with mRMR feature selection method outperformed SVM classifier. This work uses SVM and BMNB machine learning algorithms for sentiment classification. 10-fold cross validation technique is used to evaluate the results of the proposed work, since there is no separate test dataset.

Table I. Various word relation based and POS based feature sets

Features	Description
FS1	unigram features
FS2	bigram features
FS3	Trigram features
RIGFS1	Reduced unigram features with IG feature selection
RIGFS2	Reduced bigram features with IG feature selection
RIGFS3	Reduced trigram features with IG feature selection
RWFOFS1	Reduced unigram features with WFO feature selection
RWFOFS2	Reduced bigram features with WFO feature selection
RWFOFS3	Reduced trigram features with WFO feature selection
FS12	Combined feature set with unigrams and bigrams
FS13	Combined feature set with unigrams and trigrams
FS123	Combined feature set with unigrams, bigrams and trigrams
RIGFS12	Combined feature set with reduced unigrams and bigrams with IG feature selection
RIGFS13	Combined feature set with reduced unigrams and trigrams with IG feature selection
RIGFS123	Combined feature set with reduced unigrams, bigrams and trigrams with IG feature selection
RWFOFS12	Combined feature set with reduced unigrams and bigrams with WFO feature selection
RWFOFS13	Combined feature set with reduced unigrams and trigrams with WFO feature selection
RWFOFS123	Combined feature set with reduced unigrams, bigrams and trigrams with WFO feature selection
POS1	adjective + adverb
POS2	verb + noun
RIGPOS1	Reduced POS1 features with IG feature selection
RIGPOS2	Reduced POS2 features with IG feature selection
RWFOPOS1	Reduced POS1 features with WFO feature selection
RWFOPOS2	Reduced POS2 features with WFO feature selection
POS12	Combined feature set with reduced POS 1 and 2
RIGPOS12	Combined feature set with reduced POS 1 and 2 with IG feature selection
RWFOPOS12	Combined feature set with reduced POS 1 and 2 with WFO feature selection

Table II. Comparison of various basic feature extraction methods for the sample sentence "I would recommend this book because it is really interesting"

Method	Features
Unigram	[I, would, recommend, this, book, because, it, is, really, interesting]
Bigram	[I would, would recommend, recommend this, this book, book because, because it, it is, is really, really interesting]
Trigram	[I would recommend, would recommend this, recommend this book, this book because, book because it, because it is, it is really, is really interesting]
Adjective + Adverb	[interesting, really, is really, really interesting, is really interesting]

Verb+ Noun	[would, recommend, is, book, would recommend, recommend this, it is, is really, this book, book because, I would recommend, would recommend this, it is really, this book because]

IV. DATASET, EVALUATION METRICS AND RESULTS

A. Dataset used

To assess the proposed feature sets and their performance, Amazon product reviews is used. This dataset contains reviews of various domains. In this study, reviews of products like DVD, books, kitchen and electronics are used [2]. Each domain consists of 1000 positive and 1000 negative labelled reviews. Books and DVD domain consists of longer reviews in comparison to electronics and kitchen domain.

B. Evaluation metrics

To estimate the sentiment classifier's performance, metrics like Accuracy, Precision, Recall and F-measure are used. For a given category ci, the values of precision, recall, F-measure and accuracy are computed as follows.

$$\text{Precision} = \frac{\text{Documents correctly classified to category ci}}{\text{Total documents classified to category ci}} \quad (4)$$

$$\text{Recall} = \frac{\text{Documents correctly classified to category ci}}{\text{Total documents in category ci}} \quad (5)$$

$$\text{F} - \text{measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (6)$$

$$\text{Accuracy} = \frac{\text{Total number of correctly classified documents}}{\text{Total number of documents}} \quad (7)$$

C. Results and Discussion

Features extracted using the two NLP based feature extraction techniques for the example sentence "I would recommend this book because it is really exciting" is presented in Table II. Among the various experiments conducted using word relation basic features, it is found that unigrams performed well in comparison to bigrams and trigrams. The reason is that trigrams contain lot of noisy and irrelevant features which reduces the classification accuracy. The accuracy of unigrams is further enhanced by using combined features. For constructing combined features sets FS12 and FS13, 60% of unigrams and 40% of either bigrams or trigrams are used. Similarly, for constructing combined feature sets FS123, 60% of unigrams and 30% of bigrams and 10% of trigrams are used. For example, to create RWFOFS123 feature set of size 10000, 6000 features are chosen from RWFOFS1, 3000 features are chosen from RWFOFS2 and 1000 features are selected from RWFOFS3. Among the combined features, combination of unigrams+bigrams gave best accuracy as shown in table V.

Similarly, using POS basic features, adjectives and adverbs produced the best accuracy than verbs and nouns. This may be due to the reason that the adjectives and adverbs are more sentiment bearing. For constructing combined feature set POS12, 60% of POS1 and 40% of POS2 are used. For example, to create RWFOPOS12 feature set of size 1000, 600 features are chosen from RWFOPOS1 and 400 features are chosen from RWFOPOS2. From the results given in table VI, it is clear that the combined POS features of all the four categories performed well in comparison to other features.

The results also show that WFO based feature selection gave better performance with suitable values of the tuning parameter λ. Also, around 10% - 20% of the features are sufficient to classify the reviews efficiently. If large number of features is used in the document term matrix construction, then it results in a sparse matrix thereby affecting the classification accuracy. Among the term weighting schemes used, it is found that both binary and term frequency (TF) based weighting schemes performed well than the term frequency-Inverse Document Frequency (TF-IDF) schemes. The reason may be because IDF penalizes those features that are present in large number of documents.

Dependency among the features used for model construction reduces the accuracy of classifiers. Since WFO based feature selection technique is used to select more important features that are not correlated among themselves, BMNB classifier performs better than SVM classifier for sentiment analysis. Further, the performance of SVM classifier degrades as the dataset size increases.

From the experiments, it can be concluded that the performance of unigrams can be further enhanced by using other features in combination. Also, the use of feature selection also helps to achieve higher accuracy with lesser number of features.

Table III. Feature vector size for all word relation features.

Features	Book	DVD	Electronics	Kitchen
FS1	14371	14730	6999	6244
FS2	122156	115403	67764	57499
FS3	131875	126310	74974	62738
FS12	136527	130133	74763	63743
FS13	146246	141040	81973	68982
FS123	268402	256443	149737	126481
RIGFS1 and RWFOFS1	10800	12000	6000	6000
RIGFS2 and RWFOFS2	5400	6000	3000	3000
RIGFS3 and RWFOFS3	1800	2000	1000	1000
RIGFS12 and RWFOFS12	9000	10000	5000	5000
RIGFS13 and RWFOFS13	10000	11000	7000	5000
RIGFS123 and RWFOFS123	18000	20000	10000	10000

Table IV. Feature vector size for all POS based features.

Features	Book	DVD	Electronics	Kitchen
POS1	5133	5107	2542	2234
POS2	17456	17044	8758	7903
POS12	22589	22151	11300	10137

RIGPOS1 and RWFOPOS1	1200	1200	600	600
RIGPOS2 and RWFOPOS2	800	800	400	400
RIGPOS12 and RWFOPOS12	2000	2000	1000	1000

Table V. Accuracy (%) for different word relation feature sets.

Features	Book		DVD	
	BMNB	SVM	BMNB	SVM
FS1	80.8	76.7	79.1	77.8
FS2	67.2	70.2	68.8	68.9
FS3	54.6	53.4	54.6	52.7
RIGFS1	83.6	82.5	83.9	81.8
RIGFS2	71.8	71.5	75.9	76.2
RIGFS3	58.3	57.1	58.1	56.7
RWFOFS1	84.5	83.9	85.3	82.8
RWFOFS2	77.8	76.4	78.1	75.3
RWFOFS3	62.3	59.4	62.8	60.4
FS12	82.8	79.6	79.9	79.5
FS13	67.8	72.3	68.4	69.1
FS123	83.1	79.9	80.2	79.9
RIGFS12	85.4	85.1	86.8	85.4
RIGFS13	80.1	81.1	80.4	81.3
RIGFS123	86.5	86.0	87.9	86.7
RWFOFS12	87.4	86.9	88.2	87.2
RWFOFS13	85.1	83.4	86.2	85.6
RWFOFS123	88.3	88.1	88.7	88.1

Table VI. Accuracy (%) for different POS based feature sets.

Features	Book		DVD	
	BMNB	SVM	BMNB	SVM
POS1	71.2	69.4	71.8	69.9
POS2	58.4	57.3	58.9	58.1
RIGPOS1	74.2	73.1	75.6	74.2
RIGPOS2	61.5	59.8	62.1	60.3
RWFOPOS1	76.4	74.7	77.3	76.5
RWFOPOS2	62.5	61.4	64.2	63.4
POS12	79.2	78.8	80.4	79.6
RIGPOS12	85.3	85.2	87.0	86.2
RWFOPOS12	86.4	85.7	87.3	86.8

V. CONCLUSION AND FUTURE WORK

In this paper, the performance of various N gram and POS based features was examined using four datasets consisting of Amazon product reviews. Combined feature sets of unigram, bigram and trigram performs better when compared to independent features. IG and WFO feature selection methods are utilized for identifying relevant features. The performance of IG and WFO as feature selection methods are investigated and concluded that WFO performs better than IG. The reason is that WFO selects prominent features based on optimal document frequency and category ratio whereas IG can only calculate the importance of the feature. Performance of BMNB is better than SVM in terms of performance and execution time. BMNB gave best performance with reduced WFO combined features (RWFOFS123) in terms of execution time and accuracy for sentiment classification. As future work, new feature extraction and feature selection techniques may be explored as the machine learning techniques require them for effective sentiment classification.

REFERENCES

1. B. Agarwal and N. Mittal, “Prominent Feature Extraction for Sentiment Analysis”, Socio-Affective Computing, Springer International Publishing, 2016.
2. J. Blitzer, M. Dredze and F. Pereira, “Biographies, Bollywood, boom-boxes and blenders: domain adaptation for sentiment classification”, ACL, 2007.
3. Y. Dang, Y. Zhang and H. Chen, “A lexicon enhanced method for sentiment classification: an experiment on online product reviews”, IEEE Intell Syst, 2010, pp. 46–53.
4. K. Dave, S. Lawrence and PM. Pennock, “Mining the peanut gallery: opinion extraction and semantic classification of product reviews”, Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519–528.
5. G. Forman, “An extensive empirical study of feature selection metrics for text classification”, J Mach Learn Res, Vol.3, No.1, 2003, pp.1289–1305
6. K.S. Kalaivani and S. Kuppuswami, “Exploring the use of syntactic dependency features for document-level sentiment classification”, Bulletin of the Polish Academy of Sciences. Technical Sciences, Vol.67, No.2, 2019, pp. 339-347.
7. P.H. Lai, “A review on the ensemble framework for sentiment analysis”, Adv. Sci. Lett., 2015, pp. 2957–2962.
8. L.P. Hung and R. Alfred, “A Performance Comparison of Feature Extraction Methods for Sentiment Analysis”, In Advanced Topics in Intelligent Information and Database Systems, Springer International Publishing, 2017.
9. B. Liu, “Sentiment Analysis and Subjectivity”, Handbook of Natural Language Processing, 2nd edn., 2010, pp. 627–666.
10. DQ. Nguyen, T. Vu and SB. Pham, “Sentiment classification on polarity reviews: an empirical study using rating-based features”, Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis, Baltimore, 2014, pp.128–135.
11. T. O’keefe and I. Koprinska, “Feature selection and weighting methods in sentiment analysis”, Proceedings of the 14th Australasian document computing symposium, Sydney, 2009, pp.67–74.
12. A. Pak and P. Paroubek, “Text representation using dependency tree sub-graphs for sentiment analysis”, Proceedings of the 16th international conference DASFAA workshop, Vol. 6637, No. 1, 2011, pp. 323–332.
13. B. Pang and L. Lee, “Opinion mining and Sentiment analysis”, Foundations and trends in information retrieval, Vol. 2, No. 1–2. Now Publishers, 2008, pp. 1–135.
14. A. Sharma and S. Dey, “Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis”, IJCA Special Issue on Advanced Computing and Comm Technologies for HPC Applications, vol. 3, 2012, pp. 15–20.
15. S. Tan and J. Zhang, “An empirical study of sentiment analysis for Chinese documents”, Expert Syst Appl, Vol.34, No.4, 2008, pp.2622–2629.
16. S. Wang and D. Li, S. Song, Y. Wei and H. Li, “A feature selection method based on Fisher’s discriminant ratio for text sentiment classification”, Proceedings of the international conference on web information systems and mining, Shanghai, 2009, pp.88–97.
17. Z.H. Deng, K. H. Luo and H. L. Yu, “A study of supervised term Weighting Scheme for Sentiment Analysis”, Expert Systems with Applications, Vol. 41, 2014, pp.3506-3513.

AUTHORS PROFILE

K.S.Kalaivani received her M.E. degree in Computer Science and Engineering from Kongu Engineering College, Perundurai. She is currently pursuing Ph.D. from Anna University, Chennai. Her re-search interests include Sentiment Analysis and Data Mining.

Prof.S.Kuppuswami is an eminent scholar, teacher, researcher and administrator having more that 40 years of experience in India and abroad. His research interests include Software Engineering and Distributed Computing.

Dr.C.S.Kanimozhiselvi received her Ph.D from Anna University in 2011. Her research interests include Data Mining and Sentiment Analysis. Her research articles are published in national and international journals.