Enhanced Sentiment Analysis and Polarity Classification Using SentiWordNet-Vocabulary

Mr.Thavasi[1] *(Assistant Professor), Dr.P.Golda Jeyasheeli*[2] *(Professor), P.Devisri*[3] *(PG student)*

sktthavasi@mepcoeng.ac.in[1], pgolda@mepcoeng.ac.in[2], devisri.s.p@gmail.com[3]

Department Of Computer Science & Engineering,

Mepco Schlenk Engineering College, Sivakasi, TamilNadu, India

Abstract – Sentiment Analysis is used in identifying the polarity of a particular context. In recent years, with the rapid growth of social media sites and user generated reviews, ratings and recommendations have a greater impact on market growth. Sentiment analysis involves the detection of sentiment content of text using natural language processing. SentiWordNet (SWN) has been extensively used as a lexical resource widely employed by many researchers for sentiment analysis and polarity classification. Polarity classification is categorizing a piece of text into positive or negative classes. The proposed method classifies the sentences from reviews. The semantic score of subjective sentences is extracted from SWN to calculate their polarity as positive, negative based on the contextual sentence structure is then used for Support Vector Machine and classification process. And also provided the recommendation to the user. Experimental evaluation is performed on seven different benchmark dataset which includes Large Movie Review dataset. The proposed framework results in high performance when compared to other research in this field.

Keywords—Natural Language Processing, Sentiment Analysis, Feature Selection, Support Vector Machine.

1. INTRODUCTION

Sentiment Analysis (SA) is the computational treatment of opinions, sentiments and subjectivity of text. Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is applied to reviews and social media for a variety of applications, ranging from marketing to customer service.

Sentiment analysis or opinion mining aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity for a document. Sentiment analysis, or opinion mining, is concerned with the extraction of subjective or objective information present in tweets, blogs, reviews or any other similar data. Different approaches can be used to categorize the text into positive or negative classes. Sentiment analysis or polarity classification is useful in various domains such as recommender systems, market search and search engines can reverse the sentiment polarity of the text. SentiWordNet (SWN) has been used globally, in various research projects, as a lexical resource for sentiment analysis and polarity classification. It is a worthy substitute for manual labelling measures as it contains sentiment information for English language.

On the other hand, to determine the sentiment analysis or polarity classification of given text, the unsupervised algorithms utilize the information confined within the text. SentiWordNet, part-of-speech(POS) tagging and bag of words(BOW) are some of the resources and techniques used for semi-supervised or unsupervised learning. Supervised learning approaches can also be used to perform sentiment analysis. Support vector machines (SVM), Naïve Bayes, k Nearest Neighbour and decision trees are a few examples of supervised approaches used for sentiment categorization.

This research is focused on improving the sentiment classification performance by utilizing a combination of unsupervised and supervised approaches.

Collaborative filtering (CF) is a popular recommendation algorithm that bases its predictions and recommendations on the ratings or behaviour of other users in the system. In this method is that other users' opinions can be selected and aggregated in such a way as in order to provide a reasonable prediction of the active user's preference.

2. RELATED WORK

A lot of researches were conducted in the field of text categorization and sentiment analysis. In practical, labeled documents are very sparse, whereas unlabeled documents are abundance. Therefore, exploiting the unlabeled data became an

active research problem using semi-/unsupervised sentiment analysis techniques. Semi-supervised machine learning methods are usually based upon bag of words, lists of opinionated terms, or other unsupervised approaches that are used sentiment lexicons. Bhaskar et al. [5] proposed a semi-supervised method for sentiment analysis. WordNet-Affect has used to identify emotions from the preprocessed version of the retrieved data. Sentiment classification has been formed using SVM where term frequency and SentiWordNet were used to identify the vector representation of data. Chenlo & Losada [9] empirically analyzed sentence features for subjectivity and polarity categorization. Support Vector Machines (SVM) and Linear Regression classifiers are employed for this task. Good subjectivity and polarity classification performance are achieved by combining unigram or bigram features with features from sentiment lexicon.

Socher et al. [7] presented a semi-supervised method are sentiment analysis and classification. A vector pace representation is used for learn multiple word phrases. The comparison of proposed method with state-of-the-art sentiment approaches are exhibits high performance for movie review datasets. Another semi-supervised approach is introduced by Ohana and Tierney [9] for sentiment analysis and classification SentiWordNet was used in conjunction with SVM in order to achieve high performance, it has been compared with unsupervised and weakly supervised approaches. Ikeda et al. [3] proposed a semi-supervised method was sentiment classification of blogs data. The documents are more abstract form of concepts, and then, arbitrary outcomes were using these training examples. Two opinion polarity datasets are used to evaluate the proposed model. The proposed framework is used to extend for multi-class classification problems. Davidov et al. [2] presented a semi-supervised method was identification on two different datasets of tweets and product reviews in fig 4. Here, the proposed method achieved high precision, recall, and f-score without any need of domain adoption.

Huang et al. [3] proposed a framework has been automatic construction of sentiment lexicons based on constraint label propagation. Prior generic lexicon and chunk dependency information are used to determine the candidate terms. Contractual and morphological constraints are defined for sentiment terms. In the next step, constraint propagation has applied to the entire collection of sentiment terms. The proposed method achieved high sentiment classification accuracy it can be compared with state-of-the-art techniques. Experimental results are high performance of proposed method for cross-domain sentiment classification. Recent researchers have presented different strategies for the development of sentiment lexicons with prior polarity scores which can then be used for sentiment analysis

Weichselbraun et al. [5] proposed a method to enhance semantic opinion mining lexicons are identifying vague terms, extracting domain-specific context information and linking it with WordNet. Notable performance improvement is observed when SenticNet was enriched and contextualized using the proposed approach. Comparison of results has been only provided with the baseline. It is very difficult to verify the superiority of the proposed approach without state-of-the-art comparison.

In order to raise the performance of sentiment orientation detection, In the proposed method combines both machine learning and linguistics. Benchmark datasets for movie reviews are incorporated and also evaluate the proposed approach; however, state-of-the-art comparison was presented for only two other research works.

The proposed approach combines semantic web techniques with NLP. This paper lack of state-of-the-art performance and then discussion about the existing gaps that are proposed research ful fills; and then is imperative to verify the effectiveness of the proposed approach. Dragoni et al. [8] explored the application was fuzzy logic in order to compute concept polarities; it can be useful for different domains. This approach has evaluated on multi-domain dataset and compared with SVM, Naïve Bayes, and maximum entropy algorithms. The proposed method outperforms these algorithms, most for precision results.

2. THE PROPOSED APPROACH

The proposed system consists of various modules.

- Data Collection
- Tag Identification
  - ✓ POS Tagging
  - ✓ POS Tagger to SWN-Tag
- Preprocessing
- Feature Extraction

- ✓ Sentiment Score
- ✓ Generate Index Term

- Label Generation
- Recommendation.

A. ARCHITECTURE DESIGN

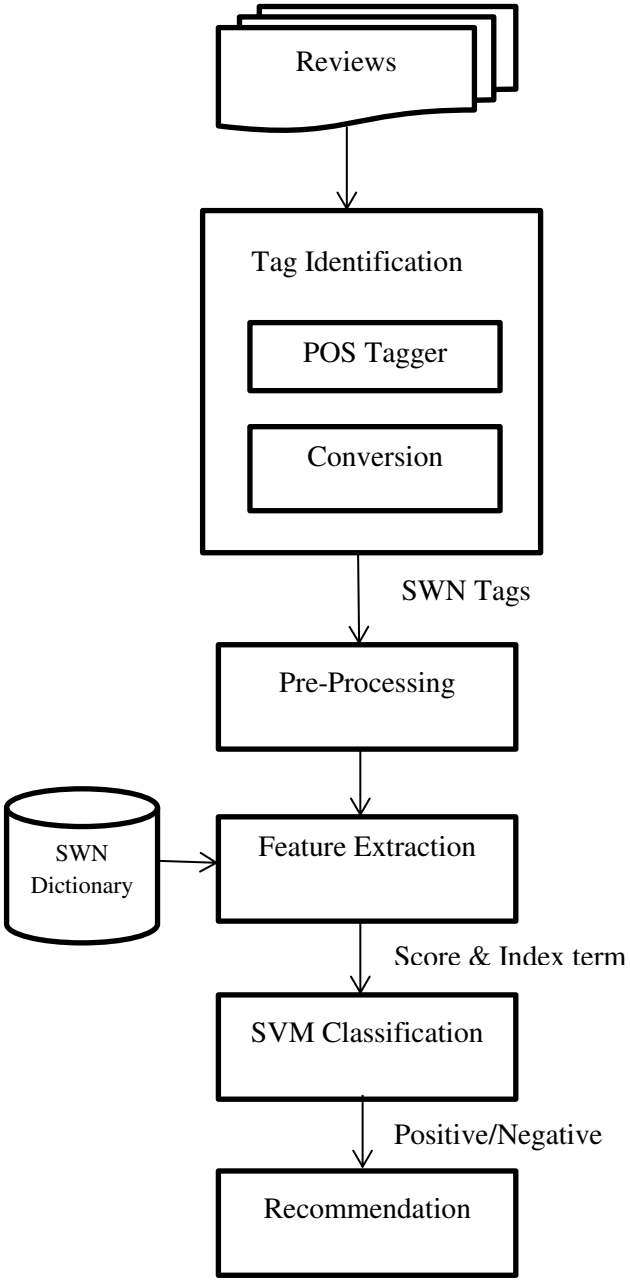

Fig.1. Architectural diagram

B. Data Collection:

Customer posts a review about the products that buy and their experience about using their product based satisfaction. The review provided by the customers is available online in social media and on the sites from where they buy the products.

The proposed approach is evaluated using seven online available benchmark datasets; Large Movie Review dataset [27]. There are a total of 50,000 movie reviews in the large movie review dataset where 25,000 are positive and 25,000 are labeled negative. It is important to note that we have used the unprocessed version of the dataset from [5], since the processed version did not include parts of speech information in fig 1.

C. Tag Identification

1) POS Tagger:

In recent research, part-of-speech information played an important role for sentiment detection. This is due to some fact that same word with different part-of-speech tag may convey a dissimilar meaning altogether. Stanford POS tagger is used in this research for tagging part of speech. It may be replaced with any other software serving the same purpose. First of all, apply the pre-processing steps which include application of POS tagging.

2) SWN Tag:

Here, transform the POS tags to SWN understandable tags and remove others.

POS	DESCRIPTION	TAG	SWN TAG
verb	Base form	VB	V
	Past tense	VBD	
noun	Singular	NN	N
	Plural	NNS	
	Proper Noun	NNP	

D. Pre-processing:

Preprocess the raw data to eliminate unnecessary data. Stop words are the most commonly occurring words in any sentence. These words are filtered prior to significance during further processing. Use stop words to remove stop words from the each review and also remove the special characters.

1) Stop Word Removal:

Stop words are another example of the most commonly used words and as they are used commonly, therefore, they

cannot be used to distinguish polarity. Some examples are, the, is, are, am, are, and, before, while, etc. Stop words are identified using a list and removed for any further processing.

- Transform the POS tag to SWN understandable tags and remove others.
- Remaining of the pre-processing steps are applied in this step, which include, stop word and special character removal etc. The input movie review is transformed as follow as fig 3.

E. Feature Selection:

Forward feature extraction method removes irrelevant features of the text and reduces original feature set. Moreover, classification accuracy is increased while decreasing the time of learning algorithm in fig 2. We have performed feature selection on movie reviews dataset after tokenization and extract subjective features identified.

1) Sentiment Score:

A general-purpose sentiment lexicon SentiWordNet3.0 (SWN) in [6]. A number of recent researchers have employed it for sentiment classification. SWN provides positivity and negativity scores for part-of-speech (POS)-tagged synsets (synonym sets). A rank is also assigned to each term in the synset based on the fact that how often that term is used in some specific sense.

Thus, a term#POS may appear multiple times with a different sense usage rank and dissimilar polarity.

Synset Score= Pos Score – Neg Score

The final Score for each term and POS pair is calculated by the equation:

$$\sum_{r=1}^{n} SynsetScore(r)/r$$

Where r is the rank of the Synset.

If the score is greater than zero, this feature is classified as positive, whereas if the score is less than zero, it is classified as negative. All the combinations with 0 scores are treated as objective and ignored for any further consideration.

- Input: Preprocessing Review, SWN dictionary
- Output: Sentimental Score

Begin

Get Pos score and Neg score

Synset score=Pos score – Neg score

For each Synset Term 1: n // n is the number of Synset term

read rank (r)

score = $\sum_{r=1}^{n} SynsetScore(r)/r$

End For

End

Fig.2. Algorithm for Sentiment Score

2) Remove Duplicate Terms:

Data duplication is a specialized data compression technique for eliminating duplicate copies of repeating data. For feature presence, all the duplicate term #POS are removed in fig 3.

3) Generate Index Term:

Now, each term #POS combination is replaced with the unique index which was generated in fig 3.

F) Label Generation:

The format is given given below. where the first column presents the class label for each instance and the next columns present the features with their respective weights.

<class_label> <feature_index>:<feature_weight> <feature_index>:<feature_weight> and so on. where

<class_label> is +1 for positive class and -1 for negative class.

<feature_index> is the unique index assigned to every feature present in the SWN-V

<feature_weight> is extracted from SWN-V based on the feature.

Finally, the review is classified as positive/Negative and results are displayed.

Input	Action	Example
"this movie must be re-released. a whole generation of comedy fans is missing out on one of the classic comedies of all time."	**POS Tagger**	"This/DT movie/NN must/MD be/VB re-released/JJ ./. a/DT whole/JJ generation/NN of/IN comedy/NN fans/NNS is/VBZ missing/VBG out/RP on/IN one/CD of/IN the/DT classic/JJ comedies/NNS of/IN all/DT time/NN ./."
	SWN Tag & Pre-processing	"movie#n be#v re-released#a whole#a generation#n comedy#n fans#n missing#v classic#a comedies#n time#n have#v never#r seen#v movie#n please#v contact#n congressman#n clergyman#n demand#n contact#v serve#v bring#v movie#n publics#n attention#n "
	Find Sentiment Score	"movie#n:0.0 be#v:0.033867874510486645 re-released#a:0.0 whole#a:0.5899908172635445 generation#n:0.0 comedy#n:0.0 fans#n:0.0 missing#v:-0.38636363636363635 classic#a:0.68comedies#n:0.0 time#n:0.0799522607053738 have#v:0.04924006245099788 never#r:-0.4583333333333333 seen#v:0.0 movie#n:0.0"
	Remove Duplicate terms	be#v:0.033867874510486645 whole#a:0.5899908172635445 missing#v:-0.38636363636363635 classic#a:0.68 time#n:0.0799522607053738 have#v:0.04924006245099788 never#r:-0.4583333333333333 0
	Generate Index Term	1269:0.033867874510486645 1270:0.5899908172635445 1271:-0.38636363636363635 1272:0.68 1273:0.0799522607053738

Fig.3. Example for proposed approach

G) Multi-Objective Model Selection (MOMS):

A novel model selection approach is proposed in this research that is based on multi objectives such as combination of precision and recall or in other words, f-measure.

Step 1: Fold Generation: Divide the datasets into 10 equal-sized mutually exclusive folds maintaining class distributions. The datasets selected for this research have equal class distributions for positive and negative instances. If any other imbalance dataset is used, it should be ensured that stratified sampling is employed.

Step 2: Training, Testing, and Evaluation Sets: The 10 folds generated in the first step are separated so that first nine folds are used for training and testing of the SVM classifier, whereas the 10th fold is used for evaluation.

Step 3: Model Generation: The data in the nine folds are again divided into 90:10 ratio, and standard 10-fold cross validation is applied. An SVM model is generated for each fold and applied to the respective test set. The precision, recall, and f-measure results for each fold of this subset are noted.

Step 4: Model Selection: Combination of precision and recall, i.e., f-measure, is compared for each result in step 3. The SVM model with the highest f-measure is selected.

Step 5: Application on Evaluation Set: The selected model is then applied on the evaluation set. The results are noted for this iteration. Steps 1–5 are repeated 10 times, each with different training, testing, and evaluation sets. The average of these results is computed.

4) Result, Evaluation and Discussion:

Accuracy, precision, recall, and f-measure are used for the evaluation of results. The equations used for these performance measures are presented as follows:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F – Measure} = 2 * \frac{Recall * Precision}{Recall + Precision}$$

where TP, TN, FP, and FN present true positives, true negatives, false positives, and false negatives respectively. Average accuracy for this approach is 88% in fig 5.

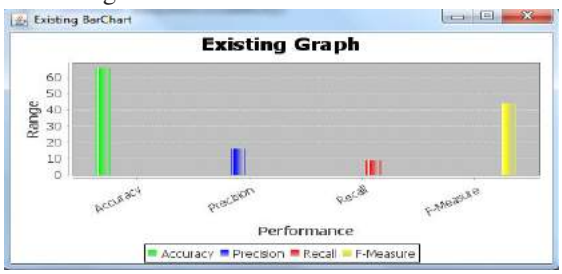

Fig 4: Accuracy for existing graph

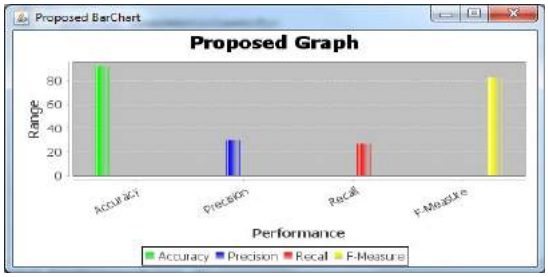

Fig 5: Accuracy for proposed graph

5) Recommendation:

Movie Recommender, a system which provides movie recommendations based on the score. The words used in user comments provide many clues about the users, about what they like or dislike and genre of the movie. The best part of the movie is predicted based on the genre of words with highest positive score. Then the movie is recommended to the users

6) Conclusion and Future work:

SentiWordNet, provide an efficient way for unsupervised text categorization. However, there is a need to improve the performance of SentiWordNet. The application of supervised learning has been the prime research focus for text classification. Labeled datasets are required in order to train supervised classifiers. This becomes the key concern as tagged datasets are not easily available and it takes huge effort and resources to build such datasets.

In future, we plan to explore other approaches like cosine similarity, information gain and gain ratio in a transfer learning methodology to further improve SentiWordNet performance.

REFERENCES

[1]Pang B, Lee L. Opinion mining and sentiment analysis. Found Trend Inf Retr. 2008;
[2] Molina-Gonza´lez MD, Martı´nez-Ca´mara E, Martı´n-Valdivia MT, Uren˜a-Lo´pez LA. A Spanish semantic orientation approach to domain adaptation for polarity classification. Inf Process Manage.2015
[3] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J. 2014;5(4): 1093–113.
[4] Saif H, He Y, Fernandez M, Alani H. Contextual semantics for sentiment analysis of Twitter. Inf Process Manag. 2015.
[5] Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl Based Syst. 2015
[6] Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexiconbased methods for sentiment analysis. Comput Ling. 2011;37(2): 267–307.
[7] Kang H, Yoo SJ, Han D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Syst Appl. 2012;39:6000–10.;
[8] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. CS224 N project report, Stanford. 2009.
[9] Kouloumpis E, Wilson T, Moore J. Twitter sentiment analysis: the good the bad and the omg! I Barcelona, Spain, 2011.