# Conceptual Sentiment Analysis Model

**Kranti Vithal Ghag[1], Ketan Shah[2]**
[1]Information Technology Department, MET's SAKEC, Mumbai University, India
[2]Information Technology Department, SVKM's NMIMS MPSTME, India

**Article Info**

***Article history:***

Received Jan 14, 2018
Revised Apr 20, 2018
Accepted Apr 27, 2018

***Keyword:***

Concept-based approach
Movie reviews
Polarity classification
Sentiment analysis
Text mining

## ABSTRACT

Bag-of-words approach is popularly used for Sentiment analysis. It maps the terms in the reviews to term-document vectors and thus disrupts the syntactic structure of sentences in the reviews. Association among the terms or the semantic structure of sentences is also not preserved. This research work focuses on classifying the sentiments by considering the syntactic and semantic structure of the sentences in the review. To improve accuracy, sentiment classifiers based on relative frequency, average frequency and term frequency inverse document frequency were proposed. To handle terms with apostrophe, preprocessing techniques were extended. To focus on opinionated contents, subjectivity extraction was performed at phrase level. Experiments were performed on Pang & Lees, Kaggle's and UCI's dataset. Classifiers were also evaluated on the UCI's Product and Restaurant dataset. Sentiment Classification accuracy improved from 67.9% for a comparable term weighing technique, DeltaTFIDF, up to 77.2% for proposed classifiers. Inception of the proposed concept based approach, subjectivity extraction and extensions to preprocessing techniques, improved the accuracy to 93.9%.

***Corresponding Author:***

Kranti Vithal Ghag,
IT Department, MET's SAKEC,
Mumbai University,
W. T. Patil Marg, Chembur, Mumbai, Maharashtra, India 400088.
Email: kranti.ghag@sakec.ac.in

## 1. INTRODUCTION

The web which is massively increasing resource of information has changed from read only to read write. Large amount of opinionated data is generated by online activities like social media, blogging, surveys and forums [1]. Electronic Word of Mouth (eWoM) has become more popular than the traditional Word of Mouth (WoM) publicity [2]. Automatically processing sentiment data needs to be handled systematically.

Sentiment Analysis involves extracting, preprocessing, understanding, classifying & presenting sentiments expressed by the users [3]. Sentiment analysis generally involves classifying the polarity of a piece of text as positive, negative or neutral. It also involves subjectivity extraction [4], intensity prediction [5] and emotion classification [6]. Sentiment analysis was also performed at the term, sentence, paragraph, document level and was also extended to aspect level [7], [8]. Sentiment Analysis was performed on languages like English, Chinese, Arabic and Vietnamese language. There are very few multilingual sentiment classifiers [9].

Sentiment analysis techniques can be broadly classified as supervised learning and unsupervised learning techniques [10]. Many unsupervised learning techniques use existing lexical resources like WordNet [11] and language specific sentiment information like sentiment seed words, their Synonyms and antonyms to construct and update sentiment lexicons [5], [12]. Unsupervised learning techniques assigned a generalized polarity and weight to a term and thus fail to capture its domain specific context. Supervised learning techniques use a set of reviews, tagged positive or negative to train the classifiers. Preprocessing tasks such as stopwords removal [13], punctuations removal are performed on these documents. Terms are then classified on the basis of their dominancy in positively tagged documents versus negatively tagged documents. This lexicon is then used for classifying the reviews using bag-of-words (BOW) approach [14].

Supervised as well as unsupervised techniques used bag-of-words approach. Processing structured vectors was systematic compared to unstructured, but positional information carried by term was ignored. Syntactic structure of a sentence plays an important role in sentiment analysis [15]. Term-document vectors provided the count of term in document but the term as an individual may carry a different meaning than what it meant as group of words in sentence. Semantic structure of words in sentence contributes to the actual meaning & thus may adjoin for sentiment analysis also. This information was not handled when representing documents as term vectors. Bag-of-words based approaches thus fail to capture syntactic and semantic structure of reviews. Although supervised techniques such as SVM, NN & unsupervised approaches such as lexicon-based approaches are popular, intelligent systems such as concept-based approaches is need of hour [1].

Sentiment Classifiers, preprocessing techniques & deduction methodologies are the core components of Supervised Sentiment Analysis. Pang, Lee, & Vaithyanathan stated that sentiment analysis needs to be handled in a more sophisticated way than traditional text categorization [16]. They are pioneers for extracting, transforming & tagging popular movie review dataset [17].Domain specific sentiment analysis models were designed for various domains such as movie, restaurant, mobile, books and DVD’s. The movie domain was relatively difficult to classify [18]. Top ranked index terms were not the top ranked sentimentally polarized terms. So terms were classified on the basis of term presence distribution across positively and negatively tagged documents [19]. Frequency distribution was not considered. Classifiers such as Naive Bayes, SVM and Random Forest Classifiers were among popular techniques for sentiment classification [20]. These classifiers classified a term as positive if it was more frequent in positively tagged documents, negative if more frequent in negatively tagged documents and neutral if equally distributed. Neutral terms might not have exactly equal occurrence in positively tagged documents and negatively tagged documents. Even if the term was distributed with a slight difference, the term was classified. Actually the term might be neutral. Classifiers also suffered due to high dimensionality if preprocessing was not performed.

As the training data had grown exponentially, dimension of the input space had also splurged. Sentiment analysis being a specialized domain of text mining benefitted after text preprocessing [21]. Pre-processing includes tokenizing, eliminating tags, stopwords removal, discarding punctuations & symbols, stemming & lemmatization [22]. Ensemble approaches were used to identify an appropriate combination of preprocessing [23]. Punctuation marks are important when writing in English or any natural language, but are hardly of any use for computational linguistics [24]. Apostrophe was discarded. The terms with apostrophe would then form a new dimension, defeating the purpose of dimensionality reduction. For example: After apostrophe removal term *didn’t* was mapped to *didnt*. The term *didnt* not being a word in English, mapped to a new dimension. There wasn’t any standard stopwords list [13]. The available stopwords list didn’t include domain-specific stopwords. Discarding neutral terms or sentiment stopwords was also a herculean task.

Training set was enriched by adding antonyms of the tagged reviews in term-document vectors of opposite orientation [25]. Negation handling was not performed. Implicit and explicit negation modifiers were handled using a three stage model [26]. The polarity shifter term might not modify the term exactly next to it. The other drawback was the availability of exact antonym. Scope of linguistic negation was determined by using Dependency Analysis [27]. Sentences were represented in form of parse trees using Stanford Parser [28]. Initial score of feature descriptors were based on SentiWordNet. 93.75% of the synonymous sets in SentiWordNet are ignored as they have a stronger objective tendency than positive and negative [29].

Processing structured vectors was easier and systematic as compared to unstructured text reviews, but the positional information carried by a term was not considered. The syntactic structure of the sentence plays an important role in sentiment analysis [15]. Term-document vectors provided the count of the term in document but the term might be associated with one or more terms in the document. This information was not recorded when representing terms as documents. The term as an individual may carry a different meaning than what it meant as group of words in the sentence structure. The semantic structures of the words in the sentence contribute to the actual meaning and thus may adjoin for sentiment analysis also. The bag-of-words based approaches failed to capture the syntactic and semantic structure of reviews.

Although supervised techniques such as SVM, NN and unsupervised approaches such as lexicon based approaches are very popular, intelligent systems such as concept based approaches is the need of the hour [1].

### 1.1. Contribution

Proposed model attempts to go beyond traditional Bags-of-words approach. It considers syntactic and the semantic structure of sentences in reviews. Terms in a review were classified as positive or negative based on their position & association with other terms in the review. The association among the terms and the position of the term in the sentence was considered. Classifiers were designed to handle neutral terms, on the basis of relative dominance. Existing preprocessing techniques were extended to handle terms with apostrophe. Rules were set to handle terms with apostrophe before discarding punctuations. A set of domain specific rules for the movie domain were proposed in the deduction phase. Subjectivity extraction was done at phrase level.

Reasearch method for proposed CSAM model is described in Section 2. Results are reported & analyzed in Section 3. Conclusion & future scope are put forth in Section 4.

## 2. PROPOSED METHOD

Figure 1 presents the proposed Conceptual Sentiment Analysis Model (CSAM).

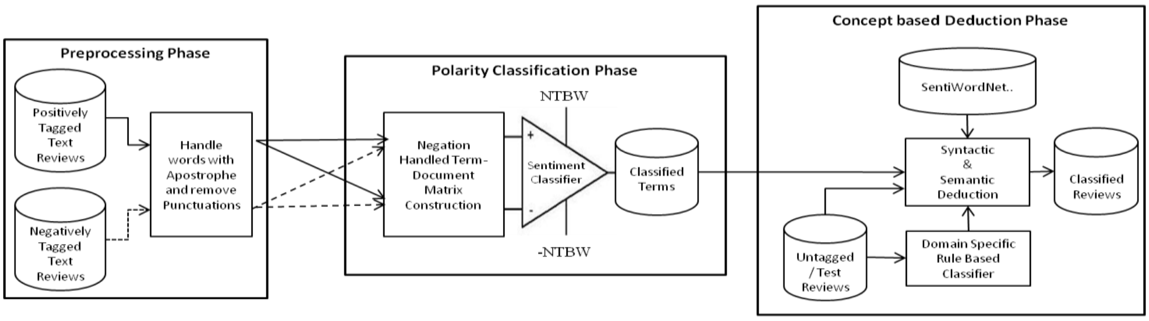
Preprocessing Phase
Positively Tagged Text Reviews
Negatively Tagged Text Reviews
Handle words with Apostrophe and remove Punctuations
Polarity Classification Phase
NTBW
Negation Handled Term-Document Matrix Construction
+
Sentiment Classifier
-
-NTBW
Classified Terms
Concept based Deduction Phase
SentiWordNet..
Syntactic & Semantic Deduction
Classified Reviews
Untagged / Test Reviews
Domain Specific Rule Based Classifier

Figure 1. Proposed conceptual sentiment analysis model

### 2.1. Preprocessing phase

The words with apostrophe such as isn’t and that’s were an overhead in term-document matrix, as illustrated with an example in Table 1. Unlike traditional preprocessing technique where punctuation symbols are discarded, a set of rules to handle words with apostrophe, are proposed in Table 2. After proposed extension of handling terms with apostrophe, traditional preprocessing task of discarding punctuations and other symbols were performed.

Table 1. Example of Term-Document Matrix entries for word “isn’t” before and after handling apostrophe

| Before Handling Terms with Apostrophe | | | | | After Handling Terms with Apostrophe | | |
|---|---|---|---|---|---|---|---|
| Documents | Terms | | | | Documents | Terms | |
| | Is | not | isn’t | | | Is | not |
| Document 1 | 1 | 3 | - | → | Document 1 | 1 | 3 |
| Document 2 | - | 1 | 1 | | Document 2 | 1 | 2 |
| Document 3 | 1 | 1 | - | | Document 3 | 1 | 1 |

Table 2. Set of rules for Handling terms with Apostrophe

| No | Rule | | | Example | | |
|---|---|---|---|---|---|---|
| 1 | n’t | → | _ not | wasn’t | - | was not |
| 2 | ’s | → | _ is | that’s | - | that is |
| 3 | ’re | → | _ are | you’re | - | you are |
| 4 | ’ve | → | _ have | they’ve | - | they have |
| 5 | ’m | → | _ am | I’m | - | I am |
| 6 | ’d | → | _would | they’d | - | they would |
| 7 | ’ll | → | _ will | you’ll | - | you will |
| 8 | ’em | → | _them | make’em | - | make them |
| 9 | in’ | → | _ ing | fringgin’ | - | frigging |

Note: Symbol “_” indicates space

### 2.2. Polarity classification phase

Preprocessed term-document matrices were provided as input to proposed Sentiment Classifier. A term was classified as positive if it was dominant in positively tagged reviews & viceversa. Dominancy of a term in reviews was determined by proposed Classifiers. Negation handling was performed at conceptual level.

Weighted Relative Term Frequency Sentiment Classifier (WRTFSC) classified term as positive if its frequency in positively tagged documents was larger than in negatively tagged documents & vice-versa

Weighted Average Relative Term Frequency Sentiment Classifier (WARTFSC) was proposed to overcome the drawback of WRTFSC. As WRTFSC was based on frequency count, it fails to handle biased data. Biased data typically includes reviews with important terms repeated too many times. To overcome this drawback, a term was classified as positive if its average frequency in positively tagged documents was larger than its average frequency in negatively tagged documents and vice-versa.

Weighted Sentiment Term Frequency Inverse Document Frequency (WSenti-TFIDF) works on the principle of relative Term Frequency Inverse Document Frequency (TFIDF) of a term across positively tagged documents and negatively tagged documents. A term was classified as positive if its TFIDF across positively tagged documents was larger than its TFIDF across negatively tagged documents and vice-versa.

Term would be classified as neutral term if its dominance in positively & negatively tagged documents was equal. Conversely even if dominance varied slightly, term would be classified as positive or negative. Ideally term should have been classified as neutral. To avoid biased classification, a window was provided defined by ± Neutral Term Window Boundary (NTWB) as shown in Figure 2 for handling neutral terms.

| Polarity$_t$ = Negative | Polarity$_t$ = Neutral | Polarity$_t$ = Positive |
|---|---|---|
| -NTWB | 0 | NTWB |

Figure 2. Sentiment Classification based on Polarity value with window for neutral words.

That is if the Polarity value of a term was between –NTWB and NTWB, the term was classified as neutral. If Polarity$_t$ value was greater than NTWB then the term was classified as positive. Conversely, if Polarity$_t$ value was lesser than −NTWB then the term was classified as negative. More words were classified as neutral if NTWB was larger, resulting to lesser number of opinionated words. Conversely if NTWB value was a smaller value neutral words would not be appropriately identified. Either condition would affect the accuracy of the classifiers. So an optimal NTWB value for each Sentiment Classification Model was experimentally determined to maximize accuracy.

Table 3. Proposed and traditional Sentiment Classifier Models

| Sr. No. | Sentiment Classifier | | Classification Criteria for term | Based on |
|---|---|---|---|---|
| 1 | TSC | Traditional Sentiment Classifier [16] | $Max(P_{ctd}, N_{ctd})$ | Frequency Count |
| 2 | Delta-TFIDF | Delta-TFIDF Sentiment Classifier [19] | $(P_{ctd} + N_{ctd}) \times \log_2\left(\frac{P_t}{N_t}\right)$ | Relative Presence Count |
| 3 | Proposed WRTFSC | Weighted Relative Term Frequency Sentiment Classifier | $\log_e\left(\frac{(P_{ctd} + 0.001)}{(N_{ctd} + 0.001)}\right)$ | Relative Frequency Count |
| 4 | Proposed WARTFSC | Weighted Average Relative Term Frequency Sentiment Classifier | $\log_e\left(\frac{\left(\frac{P_{ctd}}{P_t} + 0.001\right)}{\left(\frac{N_{ctd}}{N_t} + 0.001\right)}\right)$ | Relative Average Frequency Count |
| 5 | Proposed WSenti-TFIDF | Weighted Sentiment Term Frequency Inverse Document Frequency Sentiment Classifier | $\log_e\left(\frac{P_{ctd} \times \log_e\left(\frac{P_t}{P}\right) + 0.001}{N_{ctd} \times \log_e\left(\frac{N_t}{N}\right) + 0.001}\right)$ | Relative TFIDF values of terms |

where,

$P_{ctd}$ = Frequency of term t in positively tagged documents.
$N_{ctd}$ = Frequency of term t in negatively tagged documents.
$P_t$ = count of positively tagged documents with term t.
$N_t$ = count of negatively tagged documents with term t.
P = Total Number of positively tagged documents.
N = Total Number of negatively tagged documents.

NTWB values for all 3 proposed classifiers were experimentally determined. The classified terms were used in Concept based Deduction Phase. The mathematical models of the proposed classifier and those which were experimented for comparative analysis are summarized in Table 3.

### 2.3. Concept based deduction phase

Bag-of-words approach assumes that positions of words in reviews aren't important. It also fails to capture association within terms. Traditional approaches thus ignore syntactic & semantic structure of reviews.

Consider the reviews: Review 1: *Boring and not interesting*. Review 2: *Interesting and not boring*.

Although both of above mentioned reviews mean exactly the opposite, statistical method would treat them in the same way. Both the reviews will be represented as vectors. Each word in the review would be recorded as an element in the term-document vector with its frequency. Both vectors will be identical, though they are of different orientation. Thus statistical classifiers fail to capture their correct orientation. The proposed approaches handle the issues related to syntactic & semantic structure of sentences. Language grammar adds meaning to the text. Grammatical structure was explored using Stanford Parser [30]. It returned dependencies among terms in reviews. The proposed concept based approach then identified that negation modifier *not* is associated with term *boring* in first review thus it modifies the orientation of term *boring*. While classifying review, the polarity of term *boring* was reversed & summed with the term *interesting*, to compute the correct orientation of review 1. Similarly review 2 could also be appropriately classified. Instead of summing polarities of term in reviews, they were added by considering syntactical structure of the terms in a review.

Consider another review: A visually flashy but narratively opaque vapid exercise in style & mystification.

A classifier that follows BoW approach would sum the weights of all terms to compute review polarity. Actually the conjunction "but" notifies that only later part of sentence should be considered & prior should be ignored. Traditional approaches fail to handle these semantic relationships. Adjectives & adverbs also modify the intensity of associated terms. They should not be simply summed as in BoW approach. Semantic structure of review was systematically handled in proposed approach. Grammatical rules were explored to handle semantic structure of reviews. Stanford Parser returns dependencies among the terms in review. Polarities of the term were aggregated by considering the association of the terms in a review with other terms in the same review. Dependencies were treated to capture the correct sentiment in the review. Conjunctions, adverb modifiers and negation modifier dependencies were treated individually.

Semantic structure of example review was handled as in Figure 3. Determinant *A* was classified as stopwords. Sub-tree before conjunction *BUT* & conjunction *BUT* was discarded. Polarities of terms in adjective phrase, with adverb, *NARRATIVELY* & adjective, *OPAQUE* was multiplied. This branch was summed with noun *EXERCISE* and added to polarity of right sub-tree. In right sub-tree *IN* was classified as stopwords. Conjunction *AND* was ignored. Polarities of associated noun phrases *STYLE* & *MYSTIFICATION* were added.

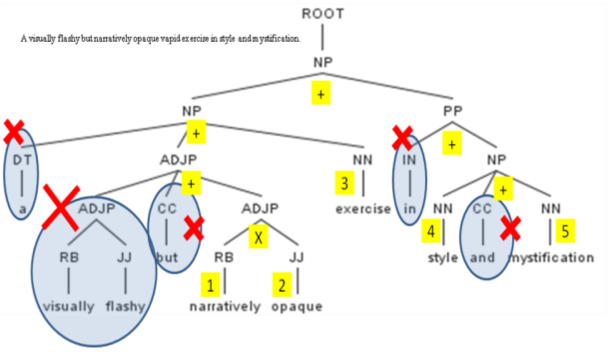

Figure 3. Grammatical dependencies for the example review using Stanford Parser

Apart from handling syntactic and semantic structure of reviews, SentiWordNet and Domain Specific Rules were also incorporated. To handle rare case of the new terms that are not present in training set, SentiWordNet was used. If a term was present in lexicon as well as SentiWordNet, the polarity of domain specific lexicon was utilized. Although only nearly 7% of the words in SentiWordNet are opinionated, sometimes these were also useful for Sentiment Classification. Polarity of the terms in constructed domain-specific lexicon was normalized to adapt with the SentiWordNet.

Domain specific rules for Movie domain were generated and expanded to improve overall Sentiment Classification accuracy. Reviewers generally remarks in a review with their final opinion about the movie. The opinion expressed about movie could be positive or negative. Two sets of rule based classifiers were designed to classify the sentiments. In first sets, various rules that are related to term Movie and all its synonyms to words with positive orientation like good, great and happy were listed. Similar terms were explored using WordNet [11]. If a review satisfied any of the condition in this it would be classified as positive. The second set was designed to classify a review as negative.

The object part of text review, which is the verb phrase of sentence, carries more opinionated information as compared to subject part or the noun phrase of the same sentence.

Consider the review. *The movie was good*.

*The movie* → subject part → informative *was good* → object part → opinionated

Considering this subjective & objective composition of a statement, subjectivity extraction was performed at phrase level. The noun phrase was ignored & the polarity for the rest of the review was computed. Noun Phrase (NP), *THE MOVIE* was ignored. The polarity was computed for the Verb Phrase (VP) *WAS GOOD*.

### 2.4. Experimental setup

Proposed Conceptual Sentiment Analysis Model (CSAM) model was evaluated on Pang & Lee's, Kaggle's & UCI's movie datasets. Proposed classifier, being domain independent, were evaluated on movie as well as restaurants and product dataset. Pang and Lee's Movie Review Dataset contains 2000 positively & negatively tagged text documents [17]. Kaggle's Bag of Words meets Bag of Popcorns dataset contains 25000 positively & negatively tagged reviews [31]. Due to computational limitation experiments were performed on Kaggle's subset of 1800 positively & negatively tagged reviews. UCI's Sentiment Labelled Movie, Restaurant & Product datasets individually consists of 1000 reviews tagged positive & negative [32].

If the label of the review and the classification outcome was same, then it contributed to accuracy. Otherwise the outcome added to error rate. Accuracy was computed using, 10 Fold Cross Validation (10 fold CV) [21]. Instead of evaluating the Conceptual Sentiment Analysis Model (CSAM) as a whole, it was evaluated in incremental manner in every experiment.

Experiment 1 was performed to determine the optimal values of Neutral Term Window Boundary (NTWB) parameter for proposed WRTFSC, WARTFSC and WSentiTFIDF classifiers. The NTWB value was varied between, 0 to ±2 in step of 0.1. Accuracy was computed at each step. Minimum value for NTWB parameter that yielded maximum accuracy was set for the respective classifiers in all further experiments.

Experiment 2 was performed to evaluate proposed classifiers. Accuracy was computed for the Traditional Sentiment Classifier (TSC), a comparable term weighting classifier (Delta-TFIDF) & proposed WRTFSC, WARTFSC and WSenti-TFIDF classifiers.

Experiment 3 was performed to evaluate proposed extensions to existing preprocessing techniques. Traditional preprocessing techniques of discarding punctuations & other symbols were applied on proposed classifiers. The proposed extension to handle terms with apostrophe was applied. Accuracy was computed for the traditional preprocessing techniques and the proposed extension to the existing preprocessing techniques.

Experiment 4 was performed to evaluate Proposed Conceptual Sentiment Analysis Model.

Experiment 5 was performed to evaluate Proposed CSAM with phrase level Subjectivity Extraction. Only the object part of the sentences in the dataset was provided as input to the CSAM model.

## 3. RESULTS AND DISCUSSIONS

Results of experiment 1, for Pang and Lee's Movie Review dataset are presented in Figure 4.

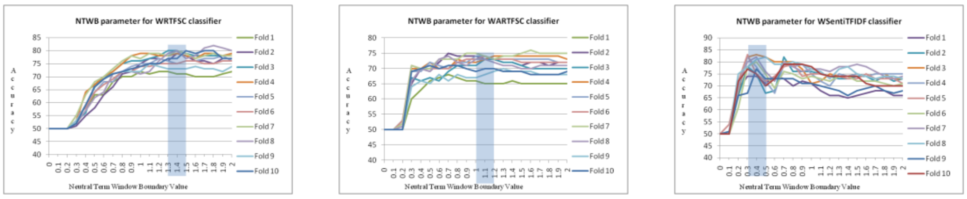

Figure 4. NTWB parameter determination for proposed WRTFSC, WARTFSC and WSentiTFIDF classifiers

Experimented NTWB values are represented on x axis. Accuracy at specific NTWB value is represented on y axis. Accuracy of each fold is represented in different colour. It can be observed from figure 4, when NTWB=0, accuracy was nearly 50%. As NTWB parameter was incremented, accuracy improved until a certain value. This NTWB value was selected as the optimal NTWB value. It can be observed from figure 4 that maximum accuracy for WRTFSC classifier was achieved at NTWB value of 1.4. Similarly it can also be noted that optimal NTWB values for WARTFSC classifier and WSentiTFIDF are 1.1 and 0.4 respectively. These are marked using a blue block. Window boundary values concept was not applicable for Traditional Sentiment Classifier (TSC) as it is not based on relative or ratio based mathematical model.

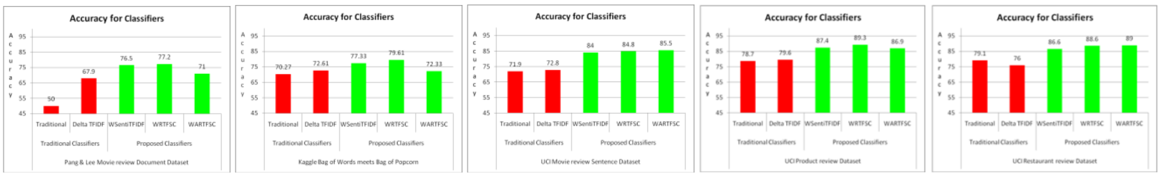

Figure 5. Accuracy Graph for evaluating Performance of Proposed Classifiers on mentioned datasets

Results for experiment 2 are presented in Figure 5. Classifiers performance was evaluated. Classifiers being independent of domain were evaluated on datasets from movie as well as product and restaurant domains. It was observed that proposed classifiers outperform the traditional classifiers. Irrespective of the input dataset the proposed classifiers achieve accuracy higher than traditional classifiers.

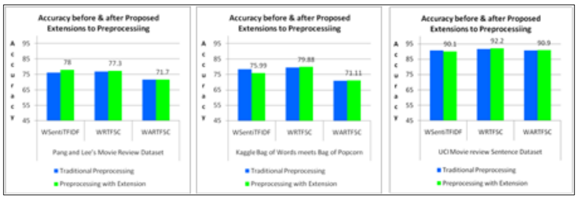

Figure 6. Accuracy graph for evaluating performance after preprocessing dataset

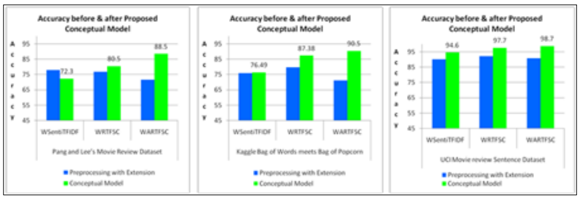

Figure 7. Accuracy graph for evaluating performance of proposed conceptual sentiment analysis model

Results for experiment 3 & 4 are presented in Figure 6 and Figure 7 respectively. A slight improvement in accuracy due to proposed extensions to existing preprocessing techniques can be observed. It can be observed that the proposed Conceptual Sentiment Analysis models (CSAM) shows remarkable improvement in accuracy.

Results for experiment 5 are presented in Figure 8. The Proposed Conceptual Sentiment Analysis model shows a remarkable improvement in accuracy after Subjectivity Extraction at the phrase level.

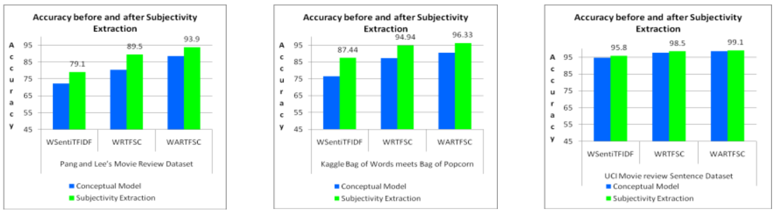

Figure 8. Accuracy graph for evaluating performance after subjectivity extraction

Results are discussed here considering the outcomes of Pang and Lee's Movie Review Dataset as it has been popularly used for analysis. Proposed classifier WRTFSC stands first with an accuracy of 77.2%. Figure 6 represents the evaluation with traditional preprocessing and proposed preprocessing extensions.

WSentiTFIDF outperformed with an accuracy of 78%. Figure 7 and Figure 8 represents the accuracy of proposed Conceptual Sentiment Analysis model and proposed Modified Conceptual Sentiment Model where WARTFSC classifier had maximum accuracy of 88.5% & 93.9% respectively.

## 4. CONCLUSION/S AND FUTURE SCOPE

Proposed WRTFSC, WARTFSC & WSentiTFIDF classifiers showed an improvement in sentiment classification accuracy from 67.9% to 77.2%.Preprocessing techniques were extended to handle term with apostrophe, along with traditional way of discarding punctuations and other symbols. A slight improvement in accuracy from 77.2% to 78% was observed due to preprocessing extensions. Bag-of-words approach failed to handle syntactic & semantic structure of reviews. As classifiers are based on probabilistic models, adapting classifiers for handling natural language specifications was an intricate task. Proposed concept based approaches handled syntactic & semantic structure of sentence. Remarkable leap in accuracy from 78% to 88.5% was observed due to proposed Conceptual Sentiment Analysis model. Subjectivity extraction was performed at the phrase level, resulting to an improvement in accuracy from 88.5% to 93.9%.

The accuracies of surveyed techniques that used Pang and Lee’s Movie Review Dataset were between 76.37% and 92.7%. Although these accuracies cannot be directly compared as the experimental setup may vary, the proposed Conceptual Sentiment Analysis Model performs better than existing techniques with a remarkable accuracy of 93.9%.

CSAM model follows the concept based approach at word, phrase and sentence level. Models that focus on the concept based approaches, at inter statement level and inter document or review level can be designed.

## REFERENCES

[1] Ravi K, Ravi V, “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications” *Knowledge-Based Systems*, 2015, 89, pp. 14-46.

[2] Nair LR, Shetty SD, Shetty SD, “Streaming Big Data Analysis for Real-Time Sentiment based Targeted Advertising, *International Journal of Electrical and Computer Engineering (IJECE),* 2017 Jan, vol. 7, no. 1, 402.

[3] Pang B, Lee LJ, Opinion mining and sentiment analysis, s.n.; 2010.

[4] K KP, S N. Insights to Problems, “Research Trend and Progress in Techniques of Sentiment Analysis”, *International Journal of Electrical and Computer Engineering (IJECE)*, 2017Jan, vol. 7, no. 5, 2018.

[5] Yu LC, Wu JL, Chang PC, Chu HS, “Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news”, *Knowledge-Based Systems*, 2013, 41, pp. 89-97.

[6] Zhang P, Wang S, Li D, “Cross-lingual sentiment classification: Similarity discovery plus training data adjustment”, *Knowledge-Based Systems*, 2016, 107, pp. 129-41.

[7] Feldman R, “Techniques & applications for sentiment analysis”, *Communications of ACM,* 2013, vol. 56, no. 4, 82.

[8] Kasthuriarachchy BH, Zoysa KD, Premaratne H, “Enhanced bag-of-words model for phrase-level sentiment analysis”, *14th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2014.

[9] Krcadinac U, Pasquier P, Jovanovic J, Devedzic V. Synesketch, “An Open Source Library for Sentence-Based Emotion Recognition”, *IEEE Transactions on Affective Computing*, 2013, vol. 4, no. 3, pp. 312-25.

[10] Ghag K, Shah K, “Comparative analysis of the techniques for Sentiment Analysis”, *2013 International Conference on Advances in Technology and Engineering (ICATE)*, 2013.

[11] Miller GA, “WordNet: a lexical database for English”, *Communications of the ACM*, 1995 Jan, vol. 38, no. 11, pp. 39-41.

[12] Neviarouskaya A, Prendinger H, Ishizuka M. SentiFul, “A Lexicon for Sentiment Analysis”, *IEEE Transactions on Affective Computing*, 2011, vol. 2, no. 1, pp. 22-36.

[13] Saini JR, Rakholia RM, “On Continent and Script-Wise Divisions-Based Statistical Measures for Stop-words Lists of International Languages”, *Procedia Computer Science*, 2016, vol. 89, pp. 313-319.

[14] Tripathy A, Agrawal A, Rath SK, “Classification of sentiment reviews using n-gram machine learning approach” *Expert Systems with Applications*, 2016, vol. 57, pp. 117-126.

[15] Socher R, “Deep Learning for Sentiment Analysis - Invited Talk”, *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2016.

[16] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?,” *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, 2002.

[17] Pang B, Lee L. A, Sentimental Education”, *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, 2004.

[18] Arora P, Virmani D, Kulkarni PS, “An Approach for Big Data to Evolve the Auspicious Information from Cross-Domains”, *International Journal of Electrical and Computer Engineering (IJECE)*, 2017 Jan, vol. 7, no. 2, 967.

[19] J Martineau J, Finin T, Joshi A, Patel S, “Improving binary classification on text problems using differential word features”, *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*. 2009.

[20] Silva N, Hruschka E, Hruschka E. Biocom Usp, “Tweet Sentiment Analysis with Adaptive Boosting Ensemble”, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014),* 2014.

[21] Han J, Kamber M, Pei J, "Data mining: concepts and techniques", Elsevier/Morgan Kaufmann; 2012.
[22] Haddi E, Liu X, Shi Y, "The Role of Text Pre-processing in Sentiment Analysis", *Procedia Computer Science*, 2013, vol. 17, pp. 26-32.
[23] Lochter JV, Zanetti RF, Reller D, Almeida TA, "Short text opinion detection using ensemble of classifiers and semantic indexing", *Expert Systems with Application*s, 2016, vol. 62, pp. 243-249.
[24] Uysal AK, Gunal S, "The impact of preprocessing on text classification", *Information Processing & Management*, 2014, vol. 50, no. 1, pp. 104-112.
[25] Xia R, Xu F, Zong C, Li Q, Qi Y, Li T, "Dual Sentiment Analysis: Considering Two Sides of One Review", *IEEE Transactions on Knowledge and Data Engineering*, 2015 Jan, vol. 27, no. 8, pp. 2120-2133.
[26] Xia R, Xu F, Yu J, Qi Y, Cambria E, "Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis", *Information Processing & Management*, 2016, vol. 52, no. 1, pp. 36-45.
[27] Padmaja S, Fatima SS, Bandu S, Sowmya B, "Comparison of the scope of negation in online news articles", *International Conference on Computing and Communication Technologies*, 2014.
[28] Indhuja K, Reghu RPC, "Fuzzy logic based sentiment analysis of product review documents", *2014 First International Conference on Computational Systems and Communications (ICCSC)*, 2014.
[29] Hung C, Lin HK, "Using Objective Words in SentiWordNet to Improve Word-of-Mouth Sentiment Classification", *IEEE Intelligent Systems*, 2013, vol. 28, no. 2, pp. 47-54.
[30] Software > Stanford Parser [Internet]. The Stanford Natural Language Processing Group. Available from: https://nlp.stanford.edu/software/lex-parser.shtml.
[31] Maas A, Daly R, Pham P, Huang D, Ng A, Potts C, "Learning word vectors for sentiment analysis", *Proceedings of the 49th Annual Meeting on Association for Computational Linguistics: Human Language Technologies - ACL '11*, 2011.
[32] Kotzias D, Denil M, Freitas ND, Smyth P, "From Group to Individual Labels Using Deep Features", *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. 2015.

## BIOGRAPHIES OF AUTHORS

**Ms. Kranti Vithal Ghag** is an Assistant Professor at Information Technology Department of MET's Shah & Anchor Kutchhi Engineering College, Mumbai, India. She is pursuing PhD in Computer Science and Engineering from SVKM's NMIMS MPSTME, Mumbai, India. Her research interests include Sentiment Analysis and Data Mining.

**Dr. Ketan Shah** has a PhD in Information Technology and is working as a Professor at SVKM's NMIMS Mukesh Patel School of Technology Management & Engineering. His research interests include Data Mining using parallel approaches and Sentiment Analysis.