International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 1 (2019) pp. 140-148

DOM_CLASSI: An Enhanced Weighting Mechanism for Domain Specific Words Using Frequency Based Probability

A. Angelpreethi[1], Dr. S. BrittoRameshKumar[2]

[1] *Assistant Professor, Department of Information Technology, St. Joseph's College Tiruchirappalli, TN, India,*

[2.] *Assistant Professor, Department of Computer Science, St. Joseph's College, Tiruchirappalli, TN, India.*

Abstract:

The information Age has created great opportunities and challenges. Information is continuously increasing day by day with flood of online reviews, blogs, chat and news. The users increasingly use social media sites to connect and share sentiments related to events and products anywhere and anytime. Twitter is one of the top most social media site. Twitter is considered as important source of information in sentiment analysis applications. Domain specific words all have the same polarity class in sentiwordnet. Therefore, identification of such words is required for sentiment classification. This paper aims to create the domain specific classifier to update the sentiment polarity and increase the classification accuracy.

Keywords: Opinion Mining, Lexicons, Tweets, Semantic Orientation.

1. INTRODUCTION

Opinion Mining (OM) and Sentiment Analysis (SA) is the application of Natural Language Processing techniques to get the subjective information from the basis text. OM focuses on discovering patterns in the text that can be analysed to classify the sentiment. Millions of People use Social Media to share their thoughts, emotion, and opinions. Social Media have received more attention nowadays. Twitter have evolved to become a great source of various kinds of information. This is due to the fact of microblogs on which people post real time messages regarding their opinions on a variety of topics. Twitter provides current issues, complain and express positive and negative sentiment for products they use in daily life.

Due to several unique characteristics of twitter, it is difficult to analyse the text using existing approaches to identify the sentiment from tweets. Therefore, it is an important task to create a unified method to mine and analyse Twitter data through the instinctive classification of tweets as positive, negative, or neutral, with prominence on a domain- specific paradigm. Sentiwordnet (SWN) is an Enhanced lexical resource for OM, which assigns to each synset of wordnet using three sentiment scores namely positive, Negative and neutral.

The proposed work is an enhancement of the previous work (angelpreethi et al 2018). The problem is the entire Domain specific words have the similar polarity class in SWN. On the other hand, their presence in labelled tweets sometimes indicates a strong association with another sentiment class. Hence, Identification of such terms are essential for accurate sentiment classification. For Example in mobile phone reviews, if a term has aggregated sentiment score is positive but its occurrence in negative reviews is sufficiently high compared to its positive occurrences. Therefore, this research aims to improve the performance of Twitter-based sentiment classification by introducing domain specific classifier.

The reminder of this article is organized as follows. In Section 2 we describe the earlier works of domain specific words classification , Section 3 describes our approach, in the in Section 4 we utilize our approach for multiple domains and finally conclusion.

2. RELATED WORK

Ivan shamshurin [1] proposed a machine learning based domain specific words extraction using SVM, naïve bayes and KNN. The experiment showed that opinion words are useful for polarity prediction but the words are not sufficient on their own and should be used only in combination with other features.

Wei Le et al [2] proposed a DWWP system which consisting of domain specific new words detection and word propagation. This paper deals with the negligence of user invented new words and converted sentiment words by means of Assembled Mutual Information.

Patricia et al[3] proposed An algorithm for expanding limited hashtags into a larger and more complete set of hashtags is proposed to collect tweets. A domain-specific sentiment lexicon is built to incorporate expressions whose sentiment varies from one domain to another, without the use of labelled data.

Masud et al. [4] proposed a lexicon- centric approach that combines different lexicons and dictionaries for the sentiment analysis of tweets. Their proposed method has different modules, namely, (i) tweet capturing and filtering, (ii) subjectivity detection, and (iii) sentiment scoring. Different lexicons, such as the opinion lexicon, WordNet, SWN, and emoticon repositories, support these modules. They achieved

92% accuracy in binary classification and more than 85% in multi- class classification. However, the system needs improvement in terms of precision in the −ive class and recall in the neutral class. Moreover, there was no facility for handling extended set of emoticons, slang, or domain- specific words in several domains.

In one more work, Prieto et al.[5] collected tweets based on different query terms, such as "flu," "depression," "pregnancy," and "eating disorders," and applied specially constructed regular expressions and machine learning algorithms for feature selection and classification to monitor public concern and disease information in Portugal and Spain. They achieved F- measure values of approximately 0.8 and 0.9, which are quite auspicious compared to the baseline methods. However, the system depends on the labelled training dataset, and there is no support for classification of emoticons, slang, and domain- dependent terms in multiple domains.

Internet slang has a strong impact on the accuracy of Twitter-based sentiment analysis applications. To address this issue, Kundi, Ahmad, Khan, and Asghar [6] presented a framework to detect and score the slang in tweets using different polarity lexicons such as SWN and other sentiment resources. They achieved better results compared with the baseline methods. The major limitations of their work include insufficient concentration on handling emoticons and the need for more sophisticated context- aware and sentiment- sensitive spell correction modules. Moreover, they did not address the issue of domain- specific words, which makes the system less effective.

Tang et al [7] proposed A propagation- centric sentiment analysis approach for Twitter. It aims at the integration of different emotional clues into a unified model and trains on both tagged and untagged datasets by switching the propagation phenomenon alternately. The experiments conducted on multiple datasets demonstrated the effectiveness of the proposed approach. The proposed method is based on general- purpose learning and can be enhanced to classify domain- specific words in different domains.

Muhamed et al[8]proposed a lexicon- enhanced polarity classification technique to compute contextual polarity at different levels. They exploited the contextual features at local and global levels. However, the SWN- based sentiment scoring technique produces incorrect polarity scores for domain-specific words. To overcome this limitation, a rich collection of domain- specific words is required to improve the performance of sentiment classification.

Tsur et al. [9] for clustering tweets, they partioned the clustering task into two distinctive tasks: batch clustering of user annotated data and online clustering of a stream of tweets.

Rangrej et al. [10] focused on clustering tweets based on its content. They compared various document-clustering techniques and showed that graph-based approach using affinity propagation performs the best in clustering short text data in the sense of clustering error.

Taboda et al [11] proposed a method is based on the analysis of text subjectivity, considering intensifiers, negations, and opinion words. They achieved better results in terms of opinionated and non-opinionated sentences. Furthermore, they created an annotated sentiment dictionary. However, their method contains no provision for analysing slang and domain-dependent words.

Aldayel and Azmi (2016) used a supervised learning technique for updating the polarity of words in reviews, whereas we compute the new sentiment score of a term using a revised scheme of polarity switching, neutral switching, and term weighting.

3. PROPOSED WORK:

From the above Literature, most of the works carried out by finding the domain specific words using a SWN based polarity. SentiWordNet (SWN) is a lexical resource, which has sufficient words and frequent updates. SWN not dependent on any domain, which has 60,000 synsets retried automatically from WordNet. SWN has three sentiment classes namely positive, Negative and Neutral Classes. The Value of each score ranges in the interval from 0.0 to 1.0 and their overall sum equals to 1 for each term.

This paper used assign accurate sentiment category and scores to domain specific words using word frequency based probability measure. There are two stages. First stage to identify the domain specific words and the next stage modification of the domain specific words using term frequency and inverse document frequency.

1. To identify the domain specific words using frequency based probability and predict the class
2. Updating a score of domain specific words using term frequency and inverse document frequency.

3.1 Sentiment Polarity class Identification:

In a Labelled reviews to find the term occurrence, which belongs to one class, compared to another using frequency based probability.

$$\text{Polarity (t)} = \begin{cases} Positive & if\ Prob\_pos > Prob\_neg \\ Negative & otherwise \end{cases}$$

Figure 1. Polarity class Predictor

Polarity predictor classifies the tweet term if polarity of positive is greater than negative it returns positive otherwise it returns negative. Here Prob_pos indicates the probability of a word occurring in positive tweets of the training set. Similarly, Prob_neg indicates the probability of a word occurring in negative tweets of the training set. The probabilities can be computed using word frequency count, which is depicted in figure 2 and 3.

$$\text{Prob\_pos} = \frac{Frequncy(t,N+)}{|N+|}$$

Figure 2. Probability for Positive

$$\text{Prob\_neg} = \frac{Frequncy(t,N-)}{|N-|}$$

Figure 3. Probability for Negative

Frequency (t,N+) is the number of word count in a positive reviews training set. Similarly Frequency (t,N-) is the number of word count in a negative reviews training set. For example the word "ride" has the neutral sentiment (or) objective sentiment class in sentiwordnet (SWN) but term polarity of the negative class Pol_neg > term polarity for positive class Pol_pos. Therefore, it indicates that there is more association that is negative.

3.2 Modifying the Tweet Polarity Score:

In this step we are going to modify the term's sentiment score when sentiwordnet based sentiment score and probability based sentiment score are not equal. Challenging task is when the word not found in sentiwordnet, classifying the word and allocate sentiment score becomes difficult. Therefore, such kind of terms needs to be updated. Accurate scoring of such domain specific words in the sentiment analysis of that reviews.

Suppose the word is not found in sentiwordnet we need to find frequency count of the following:

i) TF – Term Frequency

ii) IDF – Inverse Document Frequency

iii) Frequency based probability

Term Frequency Vs Inverse Document Frequency:

TF: Term Frequency is how frequently a term occurs in a document. Every document may differ in length. Thus, term frequency divides the document length.

$$\text{TF}=\frac{No.of.times\ term\ occurs\ in\ a\ document}{Total\ no.of.terms\ in\ a\ document}$$

IDF: Inverse Document Frequency which measures how important the term is in the document.

$$\text{IDF}=\text{Log\_e}\left(\frac{Total\ no.of\ documents}{No.of.documents\ with\ term\ t}\right)$$

Consider a dataset containing 100 words wherein the word calm appears 3 times. The term frequency (i.e., tf) for calm is then (3 / 100) = 0.03. Now, assume we have 10 million documents and the word calm appears in one thousand of these. Then, the inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 1,000) = 4. Thus, the Tf-idf weight is the product of these quantities: 0.03 * 4 = 0.12.

In addition, to use frequency based probability. Paltoglou et al used delta-scoring technique using weighted score the word Move is neutral in SWN but it is Negative in movie domain. So the polarity modifier predicted accurate class that is Negative. There are some cases to test the polarity modifier.

$$\text{Senti\_Score\_DS}=\begin{cases} PRODUCT(TF, IDF, Prob\_pos), & if\ (W \notin \boldsymbol{swn}) AND\ Prob\_pos > Prob\_neg \\ PRODUCT(TF, IDF, Prob\_neg), & If\ (W \notin \boldsymbol{swn}) AND\ Prob\_neg > Prob\_pos \end{cases}$$

Figure 5. Polarity Modifier

```
If (Word belongs to SWN)
    Use SWN classifier;
Else if (SWN polarity class! = predicted class)
    Use modified scoring method;
End if
```

Figure 4. Rules for Polarity Prediction

Overall Tweet sentiment using Dom_classi

$$= \begin{cases} \textbf{Positive } \sum_{i=1}^{n}(DSC + ASC + SSC + EC + SWN\ senti\ score) > 0 \\ \textbf{Negative } \sum_{i=1}^{n}(DSC + ASC + SSC + EC + SWN\ senti\ score) < 0 \\ \textbf{Neutral } \sum_{i=1}^{n}(DSC + ASC + SSC + EC + SWN\ senti\ score) = 0\ V \\ \sum_{i=1}^{n}(DSC + ASC + SSC + EC + SWN\ senti\ score)\ \textbf{is objective} \end{cases}$$

Figure 6. Overall Tweet sentiment using Dom_classi

To increase the classification accuracy of the domain specific words we propose to merge the product of term frequency and inverse document frequency with frequency based probability measure.

Case1:

Suppose SWN based polarity is positive but the term has more occurrences in negative tweets then we shifted to negative polarity.

Case 2:

Suppose SWN based polarity is negative but the term has more occurrences in positive tweets then we shifted to positive polarity.

Case 3:

Suppose SWN based polarity is neutral but it occurs more frequently in positive tweets then the SWN based neutral score is shifted to its equivalent positive score.

Case 4:

Suppose SWN based polarity is neutral but it occurs more frequently in negative tweets then the SWN based neutral score is shifted to its equivalent negative score.

Case 5:

Suppose word not found in SWN use Polarity modifier to find the Polarity value for the particular tweet. Therefore, the overall score of the tweet by adding Domain Specific Slang Classifier (DSC), Shortened Slang Classifier (SSC), Acronym Classifier (ASC), Emoticon Classifier and SWN based Classifier. Suppose the Word not found it is executed using Polarity modifier for Domain specific words. Obviously, SWN is a domain independent lexical resource.So the overall tweet class is classified using Dom_classi Approach, which computes the sentiment scores of all the slang terms and emoticons and opinion words.The overall working mechanism of Dom _classi Appraoch Algorithm is depicted in figure 7 and equation is figure 6.

4. RESULTS AND DISCUSSIONS:

The supervised learning algorithms separate the dataset into testing data and training data. Testing and training datasets are used to verify method effectiveness. Here 10 fold cross validation are performed Where 90% of the data are training and 10% are testing. This Dom_classi Approach is tested with Mobile phone and Electronic Appliances Laptop and kitchen dataset.

Approach: Dom_Classi

Find an accurate sentiment class and scores to domain Specific words.
Input: Labelled tweets.
Output: Updated polarity class and value
t←tweet , SWN←SentiWordNet, polarity(t) ← polarity class predictor,
DSC ←Derivative Slang Classifier, SSC←Shortened Slang Classifier, ASC←Acronym Slang Classifier.

1. Begin
2. Pre-process the tweets
3. If negation found execute opine_neg();
Identification of domain specific words using frequency based probability and class prediction
Prob_pos ← $\frac{Frequncy(t,N+)}{|N+|}$, Prob_neg ← $\frac{Frequncy(t,N-)}{|N-|}$
3. If (Prob_pos > Prob_neg) # Polarity class Identification
    Return Prob_pos;
    Else
    Return Prob_ neg;
  End if;

Modification of Sentiment scores to domain specific words
4. If (SWN score ≠ Probability Score)
    4.1 if (t ∈ SWN) # opine_lexi
        Check polarity using SWN classifier
        If (SWN score= Positive) && (polarity(t)=Negative)
        Modified Polarity = (-1)* SWN score;
        Else if (SWN score=Negative) && (polarity (t)= Positive)
        Modified Polarity = (-1)* SWN score;
        End if;
    End if;

5. IF (t ∉ SWN)&& ($AND\ Prob\_pos > Prob\_neg$)
    Return $PRODUCT(TF, IDF, Prob\_pos)$;
  Else
    Return $PRODUCT(TF, IDF, Prob\_pos)$;
  End if

6. For each t classify using Dom_classi
    If $(DSC + ASC + SSC + EC + SWN\ senti\ score) > 0$
    Return Positive;
    Else if $(DSC + ASC + SSC + EC + SWN\ senti\ score) < 0$
    Return Negative;
    Else if $(DSC + ASC + SSC + EC + SWN\ senti\ score) = 0$
    Return Neutral;
    Else
    Return Objective;
    End if;
7. Return Tweet sentiment Class
8. End

Figure 7. Algorithm for Dom_classi Approach

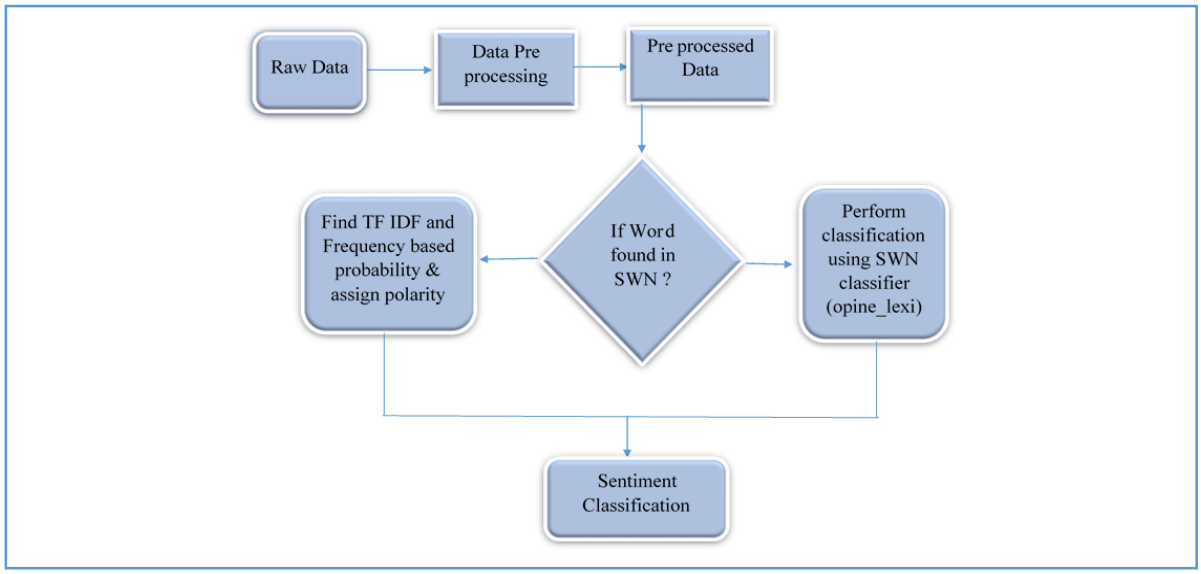

Figure 8. Methodology Diagram

Table 1. No.of Instances

Domain	Positive Instances	Negative Instances
Mobile Phone (M)	71,586	71, 654
Electronics (E)	40,879	39,846
Laptop (L)	10,315	9,918
Kitchen (K)	37,700	37,215

Table 2. Comparison of domain specific words against SWN and Dom_classi by considering Mobile as a source domain.

Domains	Domain Specific Words	No.of Words Matching With SWN
M→ E	11,805	8754
M→ L	7845	5821
M→K	6487	4957

Table 3. Comparison of domain specific words against SWN and Dom_classi by considering Electronics as a source domain.

Domains	Domain Specific Words	No. of Words Matching With SWN
E→ M	11,305	8244
E→ L	6345	4132
E→K	4479	1578

Table 4. Comparison of domain specific words against SWN and Dom_classi by considering Laptop as a source domain.

Domains	Domain Specific Words	No.of Words Matching With SWN
L→ M	14,305	11312
L→ E	8475	7894
L→K	3789	2584

Table 5. Comparison of domain specific words against SWN and Dom_classi by considering kitchen appliances as a source domain.

Domains	Domain Specific Words	No.of Words Matching With SWN
K→ M	13654	11587
K→ E	16875	13654
K→L	4798	3715

Table 6. Different domains with their Accuracy rate

Domains	M→E	M→L	M→K	E→M	E→L	E→K	L→M	L→E	L→K	K→M	K→E	K→L
Accuracy	0.8231	0.7634	0.7768	0.7534	0.7848	0.8467	0.8145	0.8710	0.8014	0.8068	0.8821	0.6998

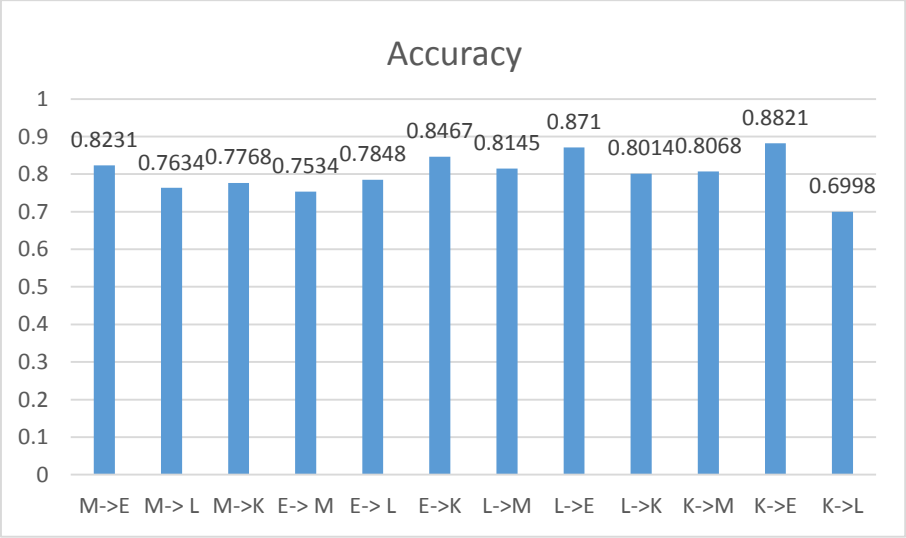

Figure 7. Accuracy comparison for different domains

M→E indicates Mobile(M) is a source and Electronics(E) is a Target domain. Similarly for all the comparisons it will take source and destinations.

Table 7. Accuracy comparison of previous works.

S.No	AUTHOR	ACCURACY (%)
1	Asghar et al	61%
2	Masud et al	81%
3	Asghar et al	83%
4	Proposed	**84.26%**

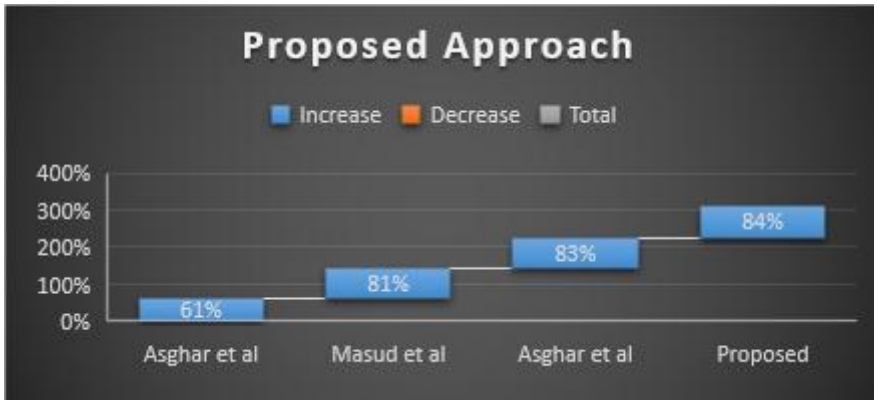

Figure 8 Comparison of Previous work Accuracy

Table 6 clearly indicates that kitchen to electronic domain accuracy rate is improved. The overall work is (opine_lexi + Dom_classi) is tested with above domains. The overall accuracy of the above work is 84.31%, which is represented in Figure 8.

5. CONCLUSION

This proposed work performed a multistage hybrid classification to perform slang, emoticon and domain specific words classification. This work is an enhancement of opine_lexi algorithm. This work can classify the tweets based on parts of speech tags and scores extracted from SWN. To increase the Accuracy of the Domain Specific Words proposed system performfrequency-based probability using term frequency and Inverse Document Frequency. The improved results in terms of accuracy shows that classification result of the proposed work is better than existing methods. In future, it will be implemented to automatic scoring of domain specific words without consulting SWN which may increase the classification Accuracy.

REFERENCES

[1] Ivan shamshurin, " Extracting domain specific opinion words for sentiment analysis",springer, 2013.

[2] LiWei, GuoKun , shi, yong, zhu luayo, zhengyuanchun, "DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain", Elsevier Knowledge based systems, Volume 146, 15 April 2018, Pages 203-214

[3] Ribeiro, P. L., Weigang, L., & Li, T. (2015). A unified approach for domain- specific tweet sentiment analysis. In Information Fusion (Fusion), 2015 18th International Conference on (pp. 846–853). IEEE.

[4] Masud, F., Khan, A., Ahmad, S., & Asghar, M. Z. (2014). Lexicon- based sentiment analysis in the social web. Journal of Basic and Applied Scientific Research, 4(6), 238–248.

[5] Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: A good place to detect health conditions. PloS One, 9(1). e86191

[6] Kundi, F. M., Ahmad, S., Khan, A., & Asghar, M. Z. (2014). Detection and scoring of internet slangs for sentiment analysis using SentiWordNet. Life Science Journal, 11(9), 66–72.

[7] Tang, J., Nobata, C., Dong, A., Chang, Y., & Liu, H. (2015). Propagation- based sentiment analysis for microblogging data. In Proceedings of the 2015 SIAM International Conference on Data Mining (pp. 577–585). Society for Industrial and Applied Mathematics.

[8] Muhammad, A., Wiratunga, N., & Lothian, R. (2016). Contextual sentiment analysis for social media genres.

Knowledge- Based Systems, 108, 92–101.

[9] O. Tsur, A. Littman, and A. Rappoport, "Efficient clustering of short messages into general domains." in ICWSM, 2013.

[10] A. Rangrej, S. Kulkarni, and A. V. Tendulkar, "Comparative study of clustering techniques for short text documents," in Proceedings of the 20th international conference companion on World wide web. ACM, 2011, pp. 111–112.

[11] A.Angelpreethi, Dr.S.Britto Ramesh Kumar," Opinion Mining on Hybrid Approach: A Perspective", International Journal of scientific research in computer science and management studies, volume 7, issue 5 (sep 2018).

[12] Alexander Pak and Patrick Paroubek, 2010. Twitter as a corpus for sentiment analysis and opinion mining. In the Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pp: 1320-1326.

[13] D.-H. Pham and A.-C. Le, "A Neural Network based Model for Determining Overall Aspect Weights in Opinion Mining and Sentiment Analysis," Indian Journal of Science and Technology, vol. 9, no. 18, 2016.

[14] A.Angelpreeethi, Dr.S.Britto Ramesh Kumar, " An Enhanced Architecture for Feature based opinion mining from product reviews", World Congress on Computing and Communication Technologies , IEEE, 2017 pp 89-92.

[15] Umar Farooq, Tej Prasad Dhamala, Antoine Nongaillard, Yacine Ouzrout and Muhammad Abdul Qadir, "A Word Sense Disambiguation Method for Feature Level Sentiment Analysis", 9th International Conference on Software, Knowledge, Information Management and Applications, IEEE, 2015.

[16] A.Angelpreethi,Dr.S.Britto Ramesh Kumar,"A Hybrid Approach To enhance the accuracy of opinion mining on bigdata using internet slang words and emoticons", International journal of Research and Analytical Reviews, volume 5, issue 4, oct-dec 2018.