MSATS: Multilingual Sentiment Analysis via Text Summarization

Rupal Bhargava*and Yashvardhan Sharma†
Department of Computer Science & Information System,
Birla Institute of Technology & Science, Pilani Campus
Pilani-333031
Email: *rupal.bhargava@pilani.bits-pilani.ac.in, †yash@pilani.bits-pilani.ac.in

***Abstract*—Sentiment Analysis has been a keen research area for past few years. Though much of the exploration that has been done supports English language only. This paper proposes a method using which one can analyze different languages to find sentiments in them and perform sentiment analysis. The method leverages different techniques of machine learning to analyze the text. Machine translation is used in the system to provide with the feature of dealing with different languages. After the machine translation, text is processed for finding the sentiments in the text. With the advent of blogs, forums and online reviews there is substantial text present on internet that can be used to analyze the sentiment about a particular subject or an object. Hence to reduce the processing it is beneficial to extract the important text present in it. So the system proposed uses text summarization process to extract important parts of text and then uses it to analyze the sentiments about the particular subject and its aspects.**

I. Introduction

Evolution in the World Wide Web has made tremendous changes in the way how human folk thinks. Earlier consultation of family members or friends were used to considered while making decisions. Likewise when any business decision was taken, associates were consulted or polling was done. With the emergence of forum, blogs, reviews etc. decisions making process has changed, for individual or business purposes. Extraction of opinions or computation of these opinions is referred to as Sentiment Analysis. Aspect based sentiment analysis analyses different subtopics or aspects present in the text that drives the original sentiment in the text. As a review, blog or forum may contain varied opinion about the subtopics. Hence it is important to analyze from such perspective. Moreover with a diversity of languages present it is important to consider different languages for sentiment analysis as well. To present such a concept, this paper proposes a technique where multiple languages are translated to the English language, to leverage the diversity of information that can be retrieved all over the world breaking the language barriers. Further, due to the existence of enormous amount of data it is best to extract the important parts of text, so that no information is left out and the redundancy and ambiguity can be removed. After this processing is complete, the text is analyzed to define the sentiments for the particular aspect present in the text thereby calculating the overall opinion. Moreover the proposed system also uses the 'user preference' if any to evaluate the overall opinion about the concerned topic. This paper is being presented as follows: Section 2 discusses the related work. Section 3 presents the problem definition which explains what the problem statement is. Section 4 explains the proposed framework and how the system is going to function cohesively. Section 5 evaluates the proposed system and discusses the results and Section 6 concludes the future work and discusses future purview.

II. Related Work

Most of the research in Subjectivity Analysis has been done in English language, and to leverage the resources present in sentiment analysis a lot of work has been done in other languages as well. Pang et al. [1] summarized the extracts by examining the alliance between subjectivity detection and polarity classification which had the same amount of polarity information as that of the full review. Mukherjee et al. [2] have used the Wikipedia knowledge to filter the irrelevant objective text and proposed a weakly supervised approach for sentiment classification of movie reviews. Ranade et al. [3] summarized online debates by extracting highly topic relevant and sentiment rich sentences. This was a done by extracting Topic relevant, document relevant and Sentiment Relevant features present in the debates. Virmani et al. [4] proposed an algorithm which clubbed aspect level with the opinion value and sentiment value to help conclude summarized value of remark about a student. Aspect tree is constructed which has different level and weights assigned to each branch to identify level of aspect. Wang et al. [5] proposed a feature based vector model and a weighing algorithm for sentiment analysis in Chinese product reviews. Also a feature extraction method based on dependency parsing is presented to identify the corresponding aspects that opinion words modify. Somprasertsri et al. [6] dedicated their work to properly identify the semantic relationships between product features and opinions. Approach presented mines product feature and opinion based on the consideration of syntactic information and semantic information by applying dependency relations and ontological knowledge with probabilistic based model. Mukherjee et al. [7] have focused mainly on the feature extraction process to identify the opinions for product review. This work is different from the proposed work by text summarization perspective.

Text Summarization has been conventionally used in sentiment analysis to filter out subjective sentences, Hu et al. [8] extracted the product features, identified the opinion sentences, and summarized the results. Their work is related to the proposed approach, but they have not considered the preferences of the user. Feczko et al. [9] had used sentiment analysis to analyze user product reviews for multiple products, identify and parse the positive and negative viewpoints and display the

978-1-5090-3519-9/17/$31.00

aggregated information in a user friendly and practically useful manner. They have not worked on the feature level, which is where the proposed approach supersedes them.

Learning is a dynamic process and to gasp the outgrowing information on internet in different languages various systems and architectures have been proposed. To achieve this, Wan [10] performed subjectivity analysis on Chinese reviews by using the English resources. Their approach first translated Chinese content to English and then combines the Chinese and English subjectivity analysis to predict the orientation. Another approach was explored by Banea et al. [11], by performing three experiments, first one being performed to obtain the subjectivity annotated sentences in Romanian from MPQA corpus, followed by using opinion finder lexicon to annotate the Romanian sentences and lastly verifying the assumption by translating in reverse direction to find new subjectivity lexicons for other language. Further they enhanced their work [12] and provided with an approach to build lexicons for language with limited resources by using boot-strapping process. A work with similar objective was done by Denecke [13], which proposed an algorithm for translating the language other than English using PROMT excellent technology and then performing sentiment analysis with the help of English resources. This is somewhat similar to the proposed work with a contrast that in proposed approach Microsoft translation API is used which supports more languages than PROMT. It is to be noted that for proposed approach Microsoft translation API was chosen instead of PROMT due to more variety of languages supported by the same. Moreover their work was based on sentence level sentiment analysis whereas the proposed algorithm supports aspect based sentiment analysis. Proposed word machine learning techniques which are better than a dictionary based approach. Apart from this has additional features such as text summarization, and user perspective which were not considered by Dencke [13]. Another work of Banea et al. [14] explored integration of features originating from multiple languages into machine learning approach for finding the subjectivity of sentence. In 2012, Balahur et al. [15] showed that it is ok to use the translation data in the context of environmental analysis. They trained the system using the translated data and performed multilingual sentiment analysis using that system. Further, they also worked [16] on tweets by collecting them and converting to small sentences. Then they translated the sentences to four different languages which were used to train the system for multilingual subjectivity analysis. An automatic framework was proposed by Banea et al [17] to explore the subjectivity information originating from multiple languages by labelling for unseen senses using either cross lingual or multilingual training with the extended bootstrapping.

III. Problem Description

Sentiment Analysis being the computational study of opinions focuses at different levels of finding the polarity of the text. There are variety of levels at which polarity of text can be classified, i.e. document level, phrase level, sentence level and feature level. As the name outlines, document level sentiment analysis classifies document being positive or negative. Likewise Sentence or phrase level sentiment analysis assumes that sentence contains only one sentiment and classifies the sentence or phrase with positive or negative polarity. But , in practical scenarios users tend to express opinions on different features or subtopics about the subject. Hence feature level sentiment analysis is one of the most important type of sentiment analysis where the polarity of each aspect of the topic is being identified. In proposed methodology sentiment polarity is being identified at feature level using text summarization. Text summarization has been used here to extract the important information from the complete text and to reduce the redundancy of data. Sentiment Analysis has been rigorously excavated for quite some time now but embedding a multilingual aspect to it will not only help everyone to work with the existing techniques but will also help acquire information from different lingual sources. Hence in this paper a method has been proposed using which user can consider multiple languages to perform sentiment analysis.

IV. Dataset

Dataset consists of documents containing reviews for 6 products(MacBook Pro, IPhone 6, IPhone 6 plus, Nexus 5, Canon ES1200D, WD hardisk) in each of the 4 languages (English, German, Spanish and French). Approx 500 reviews(per product per language) were extracted from amazon websites(amazon.in, amazon.de, amazon.es, amazon.fr) using a python script which identifies the HTML tags and retrieve the customer review content available on the website. Review extracted are stored such that reviews of one language are stored in one file for each product. Further processing and annotation has been described in Section 6 in further detail.

V. Proposed Technique

To determine the sentimental polarity of text, a pipeline process architecture is used which basically comprises of four main steps. Although these four steps are in pipeline manner but last phase of sentiment analysis is performed in parallel i.e. it calculates scores for different sentiment analysis techniques parallely. Also text summarization and aspect identification are evaluated parallely. First step mainly comprises of translation of the text to a standard language. Secondly, aspect or subtopics are extracted from the text. Thirdly, text summarization is performed to reduce the redundancy and overlapping of information present. The last phase determines the aspect based sentiment polarity, thereby calculating the overall polarity value.

A. Language Standardization

According to the given query all the documents containing the review text are retrieved irrespective of any language they're in. If all the documents retrieved contains reviews of English language system proceed further to the next phase without delaying the process but if not, the process of language identification is done for each document with different language using Microsoft API[1]. After detecting the language documents are translated to the concerned language, which in proposed case is English. All the text translated in English and already present text in English are compiled to pass to the next phase. Although Machine translators are not always perfect but in this case they can be well suited as proposed technique will not be considering the format of the text and hence the stop

[1] https://www.microsoft.com/en-us/translator/translatorapi.aspx

words etc. will not be that paramount. Prime focus of proposed approach remains with the words that actually convey some kind of sentiment and other words that are related. This can be easily achieved by the translator itself.

Algorithm 1 Language Standardization

```
Input: Documents containing reviews related to the user query.
Output: All document containing reviews in English language
Assumption: Each document is in single language.
for each document do
    if language.detect() = English then
        Continue
    else
        Translate()
    end if
end for
```

B. Aspect Identification

Before proceeding to the aspect identification step, POS tagging and stemming are done using Stanford Core NLP Tools [18]. Also stop words are removed from the text using Stanford Resources. Frequently occurring aspects are then extracted using agglomerative clustering approach where proposed approach builds a dependency graph to not only identify the aspect but also to find the relationship among the opinion and aspect which is used at a later stage in finding the opinion word. Dependency graph is build such that each word becomes a vertex in a graph and neighboring words are represented using edges. POS Tags, Sentence Id (Sentence id in a document) and Word Id (Word Id in a particular sentence) are attached with each vertex, so that it doesn't looses its context. Dependency graph nouns are considered as prominent candidate for aspects. This is the important phase, as in absence of any prior knowledge about the domain of the document there is a need to grab all the aspects that can possibly belong to that domain. Extraneous aspects can be pruned by the user if needed using the interface options. User can remove any number of features required. No restriction was imposed for the same.

C. Text Summarization

When a large corpus of data is considered where a diversity of languages exists there are high chances of redundancy and duplicity. Moreover there is a need to extract important information so that sentiment analysis considers the most important opinions from all the text. To do so proposed approach used variety of already existing scores to extract important information and imply extractive summarization [19].

- Sentence To Centroid Score: Cosine similarity of sentence vector with the centroid vector of the whole document [19].
- Cue Phrase Score: Sentences containing cue phrases such as "in summary" ,"in conclusion", "our investigation", "most important","in particular", "important", "significantly", "according to the study", "the most important", "the best" [19].

Algorithm 2 Aspect Identification

```
Input: All the reviews obtained by Algorithm 1.
Output: Aspects present in the reviews.
for all the reviews do
    Preprocess() // using Stanford Core NLP Tools identify POS tags, Stemming and Stop Word Removal
end for
for all the reviews do
    depgraph = dependency_graph()// All words are vertices and sentences are formed using edges
    Noun_List = Extract_nouns(depgraph) //This function returns nouns with there frequency
end for
for each noun in Noun_List do
    if frequency >threshold then
        Add it to aspect list
    else
        Discard it
    end if
end for
```

- Sentence position score: The position score is computed as where i is the position of the sentence. (i=1 is first sentence) [20] [19].
- Sentence inclusion of numerical data: Sentences that contain numerical data are generally important. And hence are included in the summary [20] [19].
- Uppercase word score: If a sentence contains higher no of upper case characters it is considered to be an important sentence [20] [19].
- Sentence length score: Too long or too small sentences are not useful for summary hence they are scored lower [20] [19].
- Normalized Lesk Algorithm Score: The dictionary definitions (glosses) of all meaningful words are considered and intersection operation is performed between each of these glosses and the source text itself rather than the glosses of the other words. Total number of overlap for each sentence represents the weight of the sentence in the text [21] [19].
- TF-IDF Score: Based on the term-frequency of occurrence of words [20] [19].

Above mentioned scores are used to evaluate the importance of the sentences. The overall weighted score is then used for the sorting of sentence in a descending order and hence the final summarized text is retrieved on the basis of user given threshold. Summarized text is passed to the next phase of sentiment analysis.

D. Sentiment Analysis

In this phase, aspects identified in section 5.2 and summary generated in section 5.3 are passed and aspect based opinion for each identified feature in 5.2 is predicted using SentiWordNet [22] dictionary and machine learning approaches i.e K-Nearest Neighbour, Support Vector Machine, Naive Bayes. Results for the same are shown in next section. For SentiWordNet technique, scores are assigned based on the SentiWordNet

Algorithm 3 Text Summarization

```
Input: All the reviews obtained by Algorithm 1.
Output: Summarized reviews.
for all the reviews do
    Calculate scores as mentioned in Section Text Summarization.
end for
Sort all the reviews based on their score in descending order.
Extract top sentences based on the threshold.
```

[22] and WordNet [23] dictionary. On the basis of score the aspects are assigned the opinion positive, negative or neutral whereas for Machine Learning techniques system is trained using bag of words as feature for positive, negative and neutral classes, and based on the test data it is classified in these three classes. All the techniques used are combined to perform as an ensemble classifier. Output of ensemble classifier is treated as the final sentiment corresponding to the particular aspect. After performing the sentiment classification the proposed approach obtains a complete aspect list and their corresponding opinions about individual features. The overall opinion about the topic is decided on the basis of aspect opinions. Also the system provides an option for user to input its preferences. If user provides the preferences of the aspect or feature, overall opinion is calculated on the basis of user priority else each aspect evaluated is given an equal weightage [19] i.e. each aspect has its own label and overall opinion is calculated using weights assigned by the user else every aspect is given equal weight. Overall score calculated here is a weighted sum approach where different weights are given to different aspects depending upon importance of aspects.

VI. Evaluation and Discussion

MSATS system is based on the proposed technique and has been implemented in java. To evaluate the system, experiments have been conducted using the dataset explained in section 4. The reviews collected in language other than English in dataset are translated to English language. Translated documents are again checked by language experts for the spelling error using dictionary for the errors made by machine translation. Now the content available is used for extracting the aspects of the product and summary generation. For evaluating Aspect Identification Model, human experts prepared a list of actual aspects present in the text and it was then compared with the system generated obtained aspect to evaluate its efficiency. Aspect Identification Model proved its novelty by giving Precision of 0.76 and Recall of 0.9. But the extra aspects extracted can be dealt manually as user is given a choice of selecting the aspects out of the list or deleting the extraneous aspects. Summary threshold is set to 40% of the complete content available in the experiment but the threshold limits can be changed as per user requirement at any point. As this is a user modeled system aspect selection and aspect priorities of the user are also taken as an input for the experiment. If no priorities are mentioned each aspect is given equal weightage. After the summary generation and aspect detection, aspect based sentiment analysis is done on the summarized content using SentiWordNet, K Nearest Neighbor, Support Vector Machine and Naive Bayes.

Algorithm 4 Sentiment Analysis

```
Summarized Reviews from Algorithm 3 and Aspects retrieved from Algorithm 2.
Output: Aspect with their corresponding sentiment.
Assumption: Summary=Output obtained from Algorithm 3,Aspect[] = Output obtained from Algorithm 2
Senti_Sentiment=SentiWordNet (Summary, Aspect)
KNN_ Sentiment=KNN (Summary, Aspect)
SVM_ Sentiment=SVM (Summary, Aspect)
NB_ Sentiment=NB (Summary, Aspect)
for Each Aspect do
    if Senti_ Sentiment>0 then
        Pos++
    else if Senti_ Sentiment <0 then
        Neg++
    else
        Neutral++
    end if
    if KNN_ Sentiment=positive then
        Pos++
    else if KNN_ Sentiment=negative then
        Neg++
    else
        Neutral++
    end if
    if SVM_ Sentiment=positive then
        Pos++
    else if SVM_ Sentiment=negative then
        Neg++
    else
        Neutral++
    end if
    if NB_ Sentiment=positive then
        Pos++
    else if NB_ Sentiment=negative then
        Neg++
    else
        Neutral++
    end if
    if Neutral >Pos and Neutral >Neg then
        Ensemble =Neutral
    else if Pos>Neg then
        Ensemble=Positive
    else if Neg>Pos then
        Ensemble=Negative
    else
        Ensemble=Neutral
    end if
end for
```

Results are then evaluated based on the manual annotations done. Opinions are annotated such that it contains the aspects expressed by the user and corresponding actual opinions. Data is annotated using five human annotators and actual opinion of the aspect is identified using inter annotator agreement. These annotated opinions for the features are then compared with the system generated opinion to evaluate precision, recall, and f-measure. Baseline system is taken to be the system proposed by Denecke [13]. Negated structures were not considered and translation errors were not corrected for both the systems. Though due the usage of ensemble technique instead of using SentiWordNet classification as a training data proposed system is able to rule out the ambiguity of a synset problem in the proposed approach. Movie dataset [24] was used for the training purpose for proposed system. As the translation system provides with the English document for finding out the sentiments, only the aggregated result for complete data are evaluated on each product. Proposed system rely on the accuracy of Microsoft translation system for the translation purpose which seems to justify the proposed system so far. System results evaluated are shown in the figure 1,2 and 3. It is evident from these figures that there is a significant improvement over the baseline technique. Though it should be noted that system's efficiency depends on the errors made in intermediate steps and translation errors that are subjected to text that is to be analyzed.

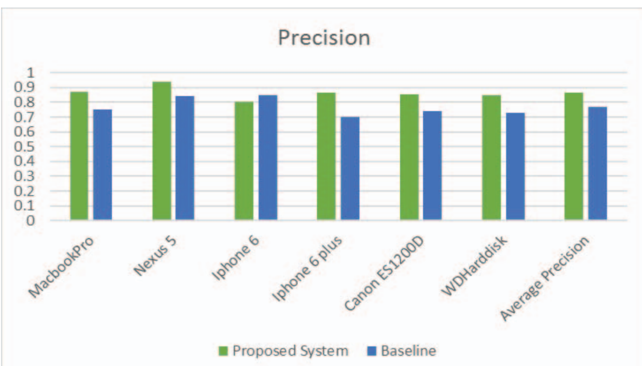

Fig. 1: Precision value for each product

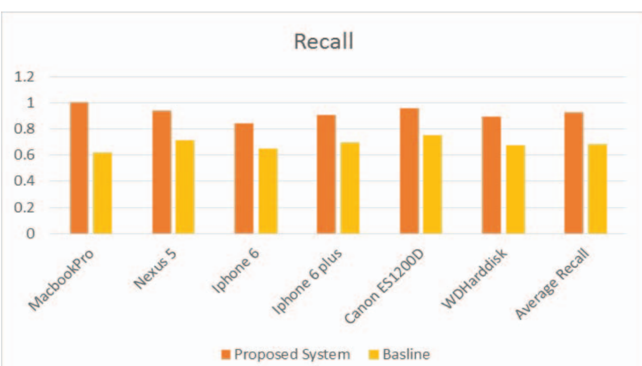

Fig. 2: Recall value for each product

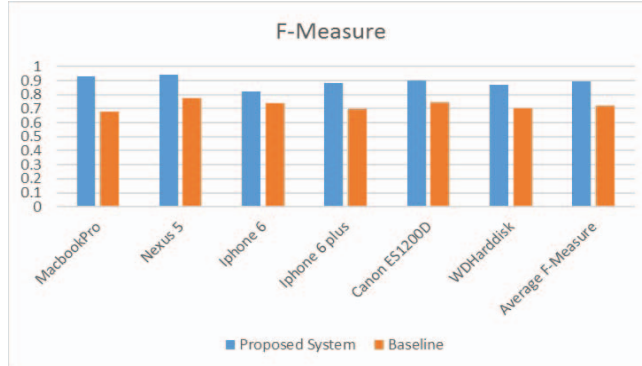

Fig. 3: F-Measure value for each product

VII. Conclusion

This paper proposes an approach for combining a variety of techniques to leverage the maximum data available of social media text and mine opinions from them. Proposed system considers multiple languages and opinions of others globally, using more than 50 languages mentioned, with a novel approach. Experiment shows that proposed technique is able to deliver promising results. Though the translator used considers over 50 languages, system can be further improved by considering regional languages like Hindi, Telugu, and Bengali etc. It will be interesting to expand the system via using linguistic features in text. Not only this, transliteration and text normalization can also be considered to further improve the system. Also technique explained for text summarization can be replaced by technique proposed in [25]. It will acquire more information in less number of sentences, pertaining to the specified threshold .

Acknowledgments

Authors would like to convey their thanks to Department of Science and Technology, New Delhi, India, for providing financial assistance under major research project scheme [S. No: SERB/F/1266/2013-2014] at the Department of Computer Science, Birla Institute of Technology and Science, Pilani, India. Authors are also grateful to the authorities of Birla Institute of Technology and Science, Pilani, for provide basic infrastructure facilities during the preparation of the paper.

References

[1] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 271.

[2] S. Mukherjee and P. Bhattacharyya, "Wikisent: Weakly supervised sentiment analysis through extractive summarization with wikipedia," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 774–793.

[3] S. Ranade, J. Gupta, V. Varma, and R. Mamidi, "Online debate summarization using topic directed sentiment analysis," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM, 2013, p. 7.

[4] D. Virmani, V. Malhotra, and R. Tyagi, "Aspect based sentiment analysis to extract meticulous opinion value," *arXiv preprint arXiv:1405.7519*, 2014.

[5] H. Wang, L. Liu, W. Song, and J. Lu, "Feature-based sentiment analysis approach for product reviews," *Journal of Software*, vol. 9, no. 2, pp. 274–279, 2014.

[6] G. Somprasertsri and P. Lalitrojwong, "Mining feature-opinion in online customer reviews for opinion summarization." *J. UCS*, vol. 16, no. 6, pp. 938–955, 2010.

[7] S. Mukherjee and P. Bhattacharyya, "Feature specific sentiment analysis for product reviews," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2012, pp. 475–487.

[8] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[9] M. Feczko, A. Schaye, M. Marcus, and A. Nenkova, "Sentisummary: Sentiment summarization for user product reviews," in *proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, 2008, pp. 265–271.

[10] X. Wan, "Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008, pp. 553–561.

[11] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan, "Multilingual subjectivity analysis using machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 127–135.

[12] C. Banea, J. M. Wiebe, and R. Mihalcea, "A bootstrapping method for building subjectivity lexicons for languages with scarce resources," 2008.

[13] K. Denecke, "Using sentiwordnet for multilingual sentiment analysis," in *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*. IEEE, 2008, pp. 507–512.

[14] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual subjectivity: Are more languages better?" in *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 2010, pp. 28–36.

[15] A. Balahur and M. Turchi, "Multilingual sentiment analysis using machine translation?" in *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*. Association for Computational Linguistics, 2012, pp. 52–60.

[16] A. Balahur and Turchi, "Improving sentiment analysis in twitter using multilingual machine translated data." in *RANLP*, 2013, pp. 49–55.

[17] C. Banea, R. Mihalcea, and J. Wiebe, "Sense-level subjectivity in a multilingual setting," *Computer Speech & Language*, vol. 28, no. 1, pp. 7–19, 2014.

[18] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit." in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[19] R. Bhargava and Y. Sharma, "User modelled aspect based semtiment analysis," in *proceedings of National Conference on Emerging Trends of Research in Applied Sciences, Experimental and Computational Techniques*. IJERT, 2015, pp. 21–24.

[20] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *Journal of emerging technologies in web intelligence*, vol. 2, no. 3, pp. 258–268, 2010.

[21] A. R. Pal and D. Saha, "An approach to automatic text summarization using wordnet," in *Advance Computing Conference (IACC), 2014 IEEE International*. IEEE, 2014, pp. 1169–1173.

[22] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." in *LREC*, vol. 10, 2010, pp. 2200–2204.

[23] G. Miller, "Wordnet-about us. wordnet. princeton university (2009)," 2010.

[24] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 142–150.

[25] R. Bhargava, Y. Sharma, and G. Sharma, "Atssi: Abstractive text summarization using sentiment infusion," *Procedia Computer Science*, vol. 89, pp. 404–411, 2016.