Literature Review on Feature Identification in Sentiment Analysis

Altaf Hussain
Capital University of Science and Technology, CUST
Islamabad, Pakistan

Shafaq Sattar
Capital University of Science and Technology, CUST
Islamabad, Pakistan

Muhammad Tanvir Afzal
Capital University of Science and Technology, CUST
Islamabad, Pakistan

ABSTRACT

Now-a-days the volume of opinions about products, issues, events, and politics etc. on different social, e-commerce and review sites grows very rapidly. From both opinion holder and opinion target point of view, it is very difficult and time consuming task to analyze all the reviews from this massive amount of data on the web. So, there is a need of efficient method that automatically extracts the opinions and relevant features of the opinion target from the reviews and finally generates the feature wise summary. Sometimes people may use different words to express same feature, this may produce a misperception in the results during feature wise summary generation. To avoid this, we need to categorize similar features for precise classification of opinions based on these feature groups. Therefore, our study is targeting the most important tasks of feature based sentiment analysis that are feature extraction and feature categorization. This paper is about to cover the currently available techniques in these two areas. We have also focused on least addressed area in this domain giving an opportunity to researchers for future work.

General Terms

Pattern Recognition

Keywords

Feature Grouping, Feature Extraction, Feature Identification, Feature Clustering, Sentiment Analysis

1. INTRODUCTION

The field of study focusing people's sentiments, opinions, evaluation, appraisals and emotions towards entities such as products, services, events, topics and their features on social web or at any online platform is known as sentiment analysis [1]. People have always been interested in what others think considering as guideline for making self decision in this regard. Social media, manufacturers and other ecommerce sites has become an easiest and publically accessible source of expressing views for a large domain of geographically distributed peoples.

Being human nature, everyone is interested to know about publically prevailed opinion regarding any burning issue or event, which may be election campaign, political affiliations, hotels rating, university grading, products review etc. In addition, manufacturer of products, owner of hotels or restaurants and political parties are as well interested to know about the public feelings towards their services or products. Social media sites are the greatest source of providing such sentiments. Moreover, there are manufacturing companies, hotel or restaurant websites maintaining an online or manual complaint system to gauge social response towards their services or products. Keeping in view these, sentiment analysis or opinion mining has become a very demanding and

fascinating field. It represents a large problem space. Sentiment analysis refers to identify and understand the feelings and attitudes of consumer towards products through the use of natural language processing (NLP). On the basis of granularity of sentiment analysis it may be divided in three levels: (i) Document level, (ii) Sentence level and (iii) Feature level sentiment analysis. Document level sentiment analysis considers the whole document as expressing opinion on a single entity. It is not applicable to documents evaluating or comparing different features of the same entity. Thus, it focuses only the polarization of whole document instead of concentrating on each feature of the opinion/sentiment target. Therefore, it does not convey the actual likes or dislikes of writer for the specific feature of a product in his/her opinionated review [2], [3]. Sentence level targets the polarization of a sentence as positive, negative or neutral. Sentence level sentiment analysis focuses on the polarization of opinionated sentences as a whole, but a writer may express multiple views for different features in a single sentence. Therefore, if writer express his likeness for battery timing but dislikes for camera result and screen display of a mobile, the analysis will mark this sentence as a neutral one due to having the same average of positive or negative phrases in sentence [2], [3]. This conveys misperception of the writer's opinion by classifying sentiment carrying sentence as neutral one. In addition, if writer express his/her opinion about likeness of the movie in just one sentence and rest of the sentences contains writer's dislikes about many features of the movie then the classification of this document with sentence level sentiment analysis will be opinionated as negative [2], [3]. Feature based sentiment analysis classify the user reviews for different features of the product as positive or negative depicting the actual sentiments of writer toward product or concerned services. Therefore, the first task in this type of sentiment analysis is features extraction from the user reviews and then applies sentiment classification for each one. It is common practice that user writes for same feature of a product or service with different wording; which may affect the actual sentiment classification at feature level as well. Therefore, we need to categorize the extracted feature prior to generate the feature based summary of opinions.

The main purpose of this review is to explore the currently available techniques on feature extraction and categorization (the steps with yellow colors in Fig. 1) in the following sections. Fig. 1 shows the general flow of feature level sentiment analysis. The rest of the paper is organized as follows. Section 2 briefly presents the recent research on feature identification. In section 3, the feature clustering is discussed. Finally conclusion is given in section 4.
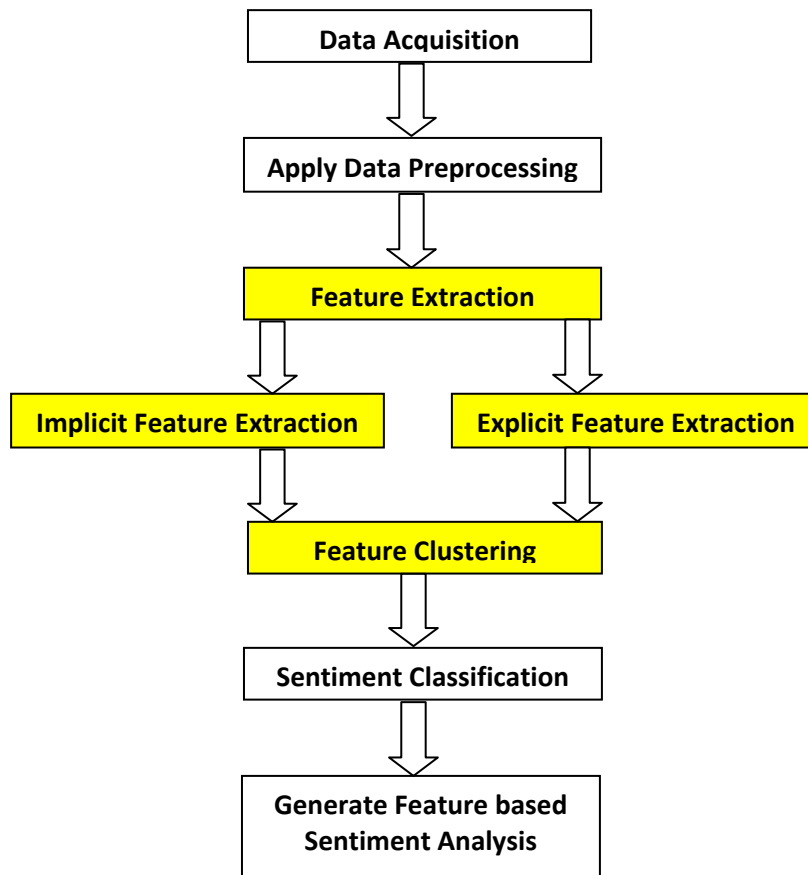
```
                    ┌─────────────────────┐
                    │  Data Acquisition   │
                    └─────────────────────┘
                              ↓
                    ┌─────────────────────┐
                    │Apply Data Preprocessing│
                    └─────────────────────┘
                              ↓
                    ┌─────────────────────┐
                    │ Feature Extraction  │
                    └─────────────────────┘
                       ↓            ↓
        ┌──────────────────────┐  ┌──────────────────────┐
        │Implicit Feature Extraction│ │Explicit Feature Extraction│
        └──────────────────────┘  └──────────────────────┘
                       ↓            ↓
                    ┌─────────────────────┐
                    │ Feature Clustering  │
                    └─────────────────────┘
                              ↓
                    ┌─────────────────────┐
                    │Sentiment Classification│
                    └─────────────────────┘
                              ↓
                    ┌─────────────────────┐
                    │ Generate Feature based│
                    │  Sentiment Analysis  │
                    └─────────────────────┘
```

Fig 1: General flow of feature sentiment analysis

2. FEATURE IDENTIFICATION

Main task in the feature level sentiment analysis is to extract relevant features from the user reviews on some entities. Feature extraction phase dealt with the extraction of explicit and implicit feature. Feature that appears directly in the review is called explicit feature e.g. "touch screen of this table is awesome". In this example "touch screen" is an explicit feature. Other type of feature that is only implied in the review is called implicit feature e.g. "This tablet will not easily fit into hand bag". In this example "fit into hand bag" indicates the feature "size". The feature "size" is only implied but it is not expressed directly. Extraction of implicit feature from implicit feature expressions can be very complex and challengeable. Feature expression is a synonym word or phrase that appears in user reviews for expressing their opinions on a specific feature of an entity. Identification of implicit features in reviews is a very complex and challengeable task; whereas literature and research work on sentiment analysis has a lot of work on explicit feature extraction.

2.1 Feature Identification using relation between opinion and its target

Literature of NLP shows that opinion words and aspects/ features of entities are always associated in some ways. Due to these relationships, opinion word can be recognized by already given or identified feature and in the same way new features can be identified by already recognized opinion words. This propagation process ends till no further opinion word and new features can be recognized. This approach assumes noun word or noun phrases as features of any service or product and assumes adjective word or adjective phrase as the opinion words that modify these features. It takes some

initial opinion lexicon as seed for initiation of extracting opinion word and features. As this process involves two ways propagation therefore the method is called Double propagation [1] also known as bootstrapping process. It is a state-of-the-art domain independent and unsupervised approach for features and sentiment words extraction process in the sentiment analysis. It eliminates the need of manual laborious time consuming work for labeling user reviews for supervised techniques.

Technique in [5] uses double propagation for extracting opinion words and features. It follows 04 extraction tasks and 03 relations for performing these extraction tasks. These tasks are: i) extracting opinion words using opinion words, ii) extracting features using opinion words, iii) extracting opinion words using features and iv) extracting features using features. For performing these tasks, the three relations used are: i) relation of opinion words and opinion words for first task, ii) relation of opinion words and features for second and third tasks, and iii) relation of features with features for fourth task. They used Minipar dependency parser to parse sentences in user reviews [5], [6]. Due to the use of seed opinion words, this method is a semi-supervised.

However, Zhang et al. [6] claimed that the work in [5] is a well suitable technique for medium-size corpora, and identified that it results in low precision and low recall for large and small corpora. In case of large corpora, double propagation results in identifying many nouns as irrelevant features and it miss some important and relevant features in case of small corpora. To overcome these two problems, Zhang et al. improved the technique by adding more relations using "Part-Whole" and "No" patterns to increase the recall. Moreover they ranked the extracted features by feature

importance that is determined by feature frequency and feature relevance. For determining feature relevance, Zhang et al. [6] uses web page ranking algorithm HITS that is a graph method. In addition this technique, with new relations, outperforms in "recall" with a minor loss in "precision" than Qui et al. [5]. The technique of relating senti-units with its target is also done in [32] using urdu text.

2.2 Feature Identification using frequent noun and noun phases

Mostly researchers have used natural language processing (NLP) techniques to extract product features from user reviews. Generally in user reviews, opinion words are adjectives or adjective phrases and product features are nouns or noun phrases. For example "touch screen of this table is awesome". In this example "touch screen" is a noun phrase as a product feature and "awesome" is an adjective as an opinion word in this sentence.

These techniques [7], [8], [9], [10], [11], [12], [13], [14], [15] and [16] of finding frequent nouns or noun phrases first apply part-of-speech (POS) tagging to all reviews, where each word in a review sentence is tagged with its part- of- speech (such as noun, adjective, adverb, verb etc.). After POS tagging it is possible to extract frequently commented nouns as product features and adjectives as opinion words. There are different freely available POS taggers like Stanford POS Tagger, SNoW-based POS tagger, tree tagger etc. But this assumption gathers too many irrelevant nouns or noun phrases and cause a noisy candidate feature list. Therefore, we need some intelligent approach to filter such noisy candidate list in feature extraction. This review finds following approaches for identifying related features among frequent nouns or noun phrases by removing irrelevant.

- Using pattern knowledge

- Using adjectives occurring in the proximity of candidate features

- Using PMI measurement between words or phrase

- Using frequency based approach; frequency-inverse sentence frequency (TS-ISF)

- Identifying pronouns referring to noun or noun phrase in candidate feature list

- Using decomposing strategy to eliminate nouns those are substrings of noun phrases in feature list

Literature related to these approaches is briefly discussed in this section as follow. Htay and Lynn [7] proposed a novel idea to find patterns knowledge in customer reviews for features and opinion extraction. They extracted nouns or noun phrases using the pattern knowledge; the Pattern knowledge defines pattern of the words that are to be extracted as features. Haty and Lynn [7] defined total eight such knowledge patterns e.g. one these knowledge patterns is: if the first word is Adjective (JJ), second word is noun (NN) and the third word is also noun then this phrase is extracted as features (high quality pictures). Using these knowledge patterns, product features are easily extracted. Similarly, Dong et al. [16] also used the pattern knowledge and consider bi-gram and single noun topics; those are automatically extracted using shallow NLP parser and statistical method. They are extracting bi-grams nouns from user reviews by using one of two basic part-of-speech co-location patterns; i) adjective followed by noun (AN) and ii) noun followed noun (NN). Bi-gram (AN) topics are filtered for being single noun topic

using the sentiment lexicon proposed in [17]. Similarly, set of nouns in reviews are identified as single noun topic but noise in these topics are removed by using double propagation technique [1]. Ojokoh et al. [11] also used patterns knowledge for providing ratings for the features to the potential customers for better online shopping.

Wang et al. [9] proposed a frequency based approach to identify the relevant noun with nearest adjective as most relevant features. This approach maintained the list of frequent candidate nouns as features after POS tagging. Then term frequency-inverse sentence frequency (TS-ISF) of the candidate terms are computed and top 20 candidates with the highest TS-ISF scores are kept in the candidate set. If the selected candidate nouns have a nearby adjective, then it is identified and picked as a frequent opinion feature from the candidate set. Finally, top five frequent opinion features are selected based on their frequency, and recommended to users [9].

Varghese and Jayasree [8] used a technique of identifying nouns or noun phrases as features after Stanford POS tagging. They also provide a co-reference resolution with Stanford Deterministic Coreference Resolution System [19], [20], [21] resolving the anaphora appearing in opinionated sentences of reviews. It relates pronouns in the following sentences to the noun or noun phrases identified as features in earlier sentences. This pronouns replacement with noun/noun phrase, prior to aspect based opinion extraction using dependency parser (Stanford Parser v1.6.4) and opinion classification using SVM classifier, is limited to only those pronouns being resolved to aspect names of the product [8].

Dalal and Zaveri [13] presents a semi-supervised approach for mining online user reviews for generating comparative feature-based statistical summaries. Parts-of-speech (POS) tagging is performed using the Link Grammar Parser (Temperley 1993) which is a well-known and efficient syntactic parser for English language (http://www.abisource.com/projects/link-grammar/). They generated a frequent feature set using multiword approach [22], [23]. A multiword (noun phrase) is a combination of multiple words that has a higher semantic significance than the individual words e.g. "face time camera", "retina display", "wireless connectivity", and "quad core graphics". Some of these multiword phrases maybe substrings of another like both the multiword "Nexus 7 front camera" and "front camera" have been extracted as frequent features. In these types of cases, decomposition strategy [24] is adopted means that more generic feature word e.g. "front camera" is extracted and discards the longer multiword. This is done because of two important reasons: (i) the opinion features to be as generic as possible, and (ii) the bloggers who post reviews are not experts and they prefer to describe products using shorter multiword features over longer ones. In addition, frequently occurring single words are also added to feature set. So in this way all noun and noun phrases are extracted as an opinion feature [13].

Wogenstein et al. [12] found the problem that sometimes POS tagging is erroneous. It happens especially when an adjective is the first word in sentence with capital letter and get tagged as noun. To deal with this problem, all such words are tagged again and the capital letter is replaced by a small letter. Stanford parser is also used and if both algorithms tag them as an adjective, then POS-tag is changed accordingly.

2.3 Feature Identification using topic modeling

Statistical topic models are used as a principle method of extracting topics from large collection of text. Topic modeling is an unsupervised approach that considers documents comprises of a mixture of topics and each of these topics is a probability distribution over words in text. The output of topic modeling is a set of word clusters that are assumed as features on extraction [1]. Related topic modeling in the domain of our review is briefly explained in this section.

Jo and Oh [33] proposed two techniques for feature extraction; first sentence-LDA and second is its extension ASUM (Aspect and Sentiment Unification Model). SLDA is used for automatic discovery of aspects/features and ASUM is used to find senti-aspect (carrying aspect and sentiment word pair). ASUM also finds some important aspects/features that are strongly related to sentiment word and did not discover by SLDA. They compared SLDA to LDA and found that LDA discovers more general and less coherent aspects. In addition, LDA targets all words regardless of sentence boundary in contrast to SLDA. But Jo and Oh (2011) claimed SLDA better in practice than LDA and assume ASUM as a technique that does not require labeled sentiments in user reviews.

Baizhang et al. [26] proposed a hybrid method that combines LDA (Latent Dirichlet Allocation) and synonym lexicon to extract features from online customer reviews on products. Input to the system is text reviews of any specific product and output is the collection of product features. They used ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System, http://www.ict.ac.cn/) to parse all sentences in Chinese customer reviews and generate POS tagging for each word. Then, noun and noun phrases are identified as initial feature set for training LDA model. Then synonym lexicons created by Mei et al. [25] is integrated with LDA to identify new features of product by adding the synonym of each term in the feature set. At the end, identified feature set are filtered by removing non-product feature terms that includes product name, time, location terms, etc. [26].

3. FEATURE CLUSTERING

Feature Clustering refers to a technique which has no predefined class labels but groups features using the similarity measures between them. It puts most similar features into one class and dissimilar into another class of features. Sometimes people tend to express the same feature with different terms e.g. "battery life", "battery" and "battery power". So clustering algorithm should assign same class to these features because it represents same aspect of a product. More specifically, in a supermarket the items like "butter", "cheese", "yogurt", and "milk" are clustered into the category of dairy products. This is a qualitative kind of classification. A qualitative approach would be to measure certain product features, for instance, products with high percent of butter would be grouped together. In general, we have "n" data points that have to be classified into "k" clusters. Simply, the purpose of clustering is to assign a cluster to each data point.

Our review finds following techniques (categorizes as semi-supervised, supervised and unsupervised approaches) that have been used for feature clustering by different researchers in their work.

3.1 Semi-supervised Approaches

Zhai et al. [27] proposed a semi-supervised approach, known as, Soft Constraints Expectation Maximization (SC-EM) for grouping feature expressions into same feature group. Feature group is the name of a feature, while a feature expression of a feature is a word or a phrase that actually appears in a review to indicate the feature. For example, a feature group could be named "picture quality", that may have many possible feature expressions e.g., "picture", "photo", "image", and even the "picture quality" itself in the user reviews indicating the same feature group. Zhai et al. [27] show that some prior knowledge can help in providing a better initialization, and consequently generate better grouping results. They produced a data set based on two pieces of natural language knowledge: (i) Feature expressions sharing some common words e.g. "battery life" and "battery power". (ii) Feature expressions that are synonyms in a dictionary e.g., "movie" and "picture". These two pieces of knowledge are named as "Soft Constraints" that help in grouping feature expressions into same feature group. Zhai et al. [27] used EM algorithm with soft constraints based on Bayesian classification for grouping feature expressions.

However, Zhai et al. [28] proposed another semi-supervised method known as constrained-LDA [24] that does not need any pre-labeled data as initial seed. Constrained-LDA is an extension of LDA method with the ability to consider existing knowledge in the form of must-link and cannot-link constraints. Latent Dirichlet Allocation (LDA) is a popular generative model to perform topic modeling. In LDA, the data is in the form of a collection of documents and each document is considered as a collection of words. Each document is assumed as mixture of latent topics and each topic is a mixture of words. A "must-link" constraint specifies that two data instances must be in the same feature cluster and "cannot-link" constraint specifies that two data instances cannot be in the same feature cluster. Two observations are used to extract "must-link" and "cannot-link" constraints automatically by this proposed technique. First, it is observed that a review sentence may include several product features that cannot be feature synonyms. It is because user did not repeat same feature in a single sentence. This observation allows us to extract many "cannot-link" constraints automatically. Secondly for "must-links", it can be observed that two noun phrases that shared one or more words are likely to fall into the same topic, e.g., "battery life", "battery consumption" and "battery power" [28].

3.2 Unsupervised Approaches

In NLP, Latent semantic analysis is a technique of analyzing semantic relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. Latent semantic analysis (LaSA) assumes that words that are similar in meaning will occur in similar group of text.

Guo et al. [29] proposed an unsupervised product-feature categorization technique with multilevel latent semantic association. Using unlabeled review corpus and the list of product features, first LaSA model capture latent semantic association among words of product features. This is done in the form of defining latent topic structure for each of these latent semantic associated words, under same topic, by assigning them different topic numbers. For Example the product feature terms are "memory card slot" and the corresponding latent topic structure for this feature is "topic20 topic4 topic4" because word "memory" is assigned to topic20 and words "card" and "slot" is assigned to the same topic4. After that, second LaSA is used to categorize the product features on the basis of their latent topic structure. The benefit of this technique is that it is language and domain independent [29]. Unlike unsupervised, supervised technique works better

for feature categorization but it requires a lot of human efforts for labeling training data.

Please use a 9-point Times Roman font, or other Roman font with serifs, as close as possible in appearance to Times Roman in which these guidelines have been set. The goal is to have a 9-point text, as you see here. Please use sans-serif or non-proportional fonts only for special purposes, such as distinguishing source code text. If Times Roman is not available, try the font named Computer Modern Roman. On a Macintosh, use the font named Times. Right margins should be justified, not ragged.

3.3 Supervised Approaches

K-means clustering is a machine learning algorithm used to cluster 'n' observations into 'k' groups of related observations, where 'k' is provided as an input parameter. First of all, the algorithm randomly selects 'k' points as the initial cluster centers ("means"). Then each data point in the dataset is assigned to the closed cluster that is based on the Euclidean distance between each data point and each cluster center Liu et al. [30] used K-means to cluster the features into groups on the basis of their high similarity.

This method perform clustering on the basis of three observations: First it considers the similarity of opinion words (means similarity of opinion words can be helpful to guide the clustering), Second it considers the similarity of features in reviews (this aspect considers the features which include the same word e.g. "running speed" and "speed" both represent the same feature "speed"), and third it considers the structure of the feature in the comment (means consider the type of feature that whether it is N (noun), NV(noun + verb), V(verb),VN(verb + noun) or NN(noun + noun) and the other is the location of feature).

4. CONCLUSION

Feature extraction is an important part of feature based sentiment analysis. Mostly researchers have used natural language processing (NLP) techniques to extract product features from customer reviews. Meanwhile, different topic modeling and supervised techniques are also used by researchers for this purpose, but it needs some training data from user. This review found a lot of research work on the feature extraction; especially on explicit feature extraction. But we observed comparatively lesser work on implicit feature extraction and feature clustering in the domain of feature based sentiment analysis. NLP is an intelligent technique to extract explicitly expressed product features from customer reviews but it cannot perform well while extracting implicitly expressed features that may have more significance in this field. In extracting implicit features the list of feature indicators are needed to be given as input [18] e.g. "expensive" and "cheap" can be used as feature indicator for "price" feature. Meanwhile, mostly techniques only consider frequent features as an important product features and ignores infrequent feature. Sometimes infrequent features may have significance according to the relevant domain. For future point of view, researchers should focus on extracting implicit features from customer reviews and also focus on the extraction of infrequent but important features.

Feature clustering is critical task and important in generating overall product summary. But limited work has been done on feature clustering. Constrained clustering techniques like constrained-LDA requires set of constraints as input and SC-EM (soft constrained- expectation maximization) requires labeled data in the form of feature expression as input whereas distance based clustering techniques like k-means clustering and fuzzy c-mean clustering requires specified number of clusters as an input. Techniques of data clustering that is based on semantic association like LSA assumes that the words that are similar will occur in same data cluster and cannot be able to detect the word that has multiple meanings. Like extraction of implicit features, feature clustering also got lesser attention in the research work. Usually synonyms in thesaurus dictionaries are used to categorize features. But products and services are commented for same features with different wording. Synonyms may have two types of ambiguities: (i) Some words may not be synonyms in thesaurus dictionaries but may refer to same feature in an application domain e.g. "photo" and "video", and (ii) Many synonyms are domain dependent e.g. movie and picture in movie review but not in camera reviews [31]. These problems are usually resolved using sharing words and lexical similarity in research. In addition some features may be assigned to different feature groups in different domains. These are some issues to be considered in feature clustering field.

5. REFERENCES

[1] Liu, B. 2012. Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies). *Morgan & Claypool Publishers.*

[2] Rashid, A., Anwer, N., Iqbal, M., and Sher, M. 2013. A survey paper: areas, techniques and challenges of opinion mining. *IJCSI International Journal of Computer Science Issues*, *10*(2), 18-31.

[3] Khushboo, T. N., Vekariya, S. K., & Mishra, S. 2012. Mining of Sentence Level Opinion Using Supervised Term Weighted Approach of Naïve Bayesian Algorithm. *International Journal of Computer Technology and Applications*, *3*(3).

[4] Motwani, M., and Tiwari, A. 2014. Comparison of Novel Semi supervised Text classification using BPNN by Active search with KNN Algorithm. *International Journal on Computer Science and Engineering*, *6*(5), 181.

[5] Qiu, G., Liu, B., Bu, J., and Chen, C. 2009. Expanding Domain Sentiment Lexicon through Double Propagation. In *IJCAI* (Vol. 9, pp. 1199-1204).

[6] Zhang, L., Liu, B., Lim, S. H., and O'Brien-Strain, E. 2010. Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 1462-1470). Association for Computational Linguistics.

[7] Htay, S. S., and Lynn, K. T. 2013. Extracting Product Features and Opinion Words Using Pattern Knowledge in Customer Reviews. *The Scientific World Journal*, *2013*.

[8] Varghese, R., and Jayasree, M. 2013. Aspect based Sentiment Analysis using support vector machine classifier. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*(pp. 1581-1586). IEEE.

[9] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[10] Kang, D., and Park, Y. 2014. Review-based measurement of customer satisfaction in mobile service:

Sentiment analysis and VIKOR approach. *Expert Systems with Applications*, *41*(4), 1041-1050.

[11] Bolanle Adefowoke Ojokoh, Olufunke Catherine Olayemi, and Olumide Sunday Adewale. 2013. Generating Recommendation Status of Electronic Products from Online Reviews. *Intelligent Control & Automation* (2153-0653) 4.1

[12] Wogenstein, F., Drescher, J., Reinel, D., Rill, S., and Scheidt, J. 2013. Evaluation of an algorithm for aspect-based opinion mining using a lexicon-based approach. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining* (p. 5). ACM.

[13] Dalal, M. K., & Zaveri, M. A. 2013. Semisupervised learning based opinion summarization and classification for online product reviews. *Applied Computational Intelligence and Soft Computing*, *2013*, 10.

[14] Anwer, N., Rashid, A., & Hassan, S. 2010. Feature based opinion mining of online free format customer reviews using frequency distribution and Bayesian statistics. In *Networked Computing and Advanced Information Management (NCM), 2010 Sixth International Conference on* (pp. 57-62). IEEE.

[15] Patel, Falguni N., and Neha R. Soni. 2012. Text mining: A Brief survey. *International Journal of Advanced Computer Research* 2.4.

[16] Dong, R., Schaal, M., O'Mahony, M. P., and Smyth, B. 2013. Topic extraction from online reviews for classification and recommendation. In*Proceedings of the Twenty-Third international joint conference on Artificial Intelligence* (pp. 1310-1316). AAAI Press.

[17] Hu, M., and Liu, B. 2004. Mining opinion features in customer reviews. In*AAAI* (Vol. 4, No. 4, pp. 755-760).

[18] Siqueira, H., and Barros, F. 2010. A feature extraction process for sentiment analysis of opinions on services. In *Proceedings of International Workshop on Web and Text Intelligence*.

[19] Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, *39*(4), 885-916.

[20] Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 28-34). Association for Computational Linguistics.

[21] Sleator, D. D., and Temperley, D. 1995. Parsing English with a link grammar.*arXiv preprint cmp-lg/9508004*.

[22] Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22-29.

[23] Zhang, W., Yoshida, T., and Tang, X. 2008. Tfidf, lsi and multi-word in information retrieval and text categorization. In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on* (pp. 108-113). IEEE.

[24] Zhang, W., Yoshida, T., and Tang, X. 2007. Text classification using multi-word features. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on* (pp. 3519-3524). IEEE.

[25] Mei, J., Lan, Y. and Gao, Y. 1983. Synonyms lexicon. *1st Edition. Shanghai Dictionary Publishing House, Shanghai.*

[26] Ma, B., Zhang, D., Yan, Z., and Kim, T. 2013. An Lda and Synonym Lexicon Based Approach to Product Feature Extraction from Online Consumer Product Reviews. *Journal of Electronic Commerce Research*, *14*(4), 304-314.

[27] Zhai, Z., Liu, B., Xu, H., and Jia, P. 2010. Grouping product features using semi-supervised learning with soft-constraints. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 1272-1280). Association for Computational Linguistics.

[28] Zhai, Z., Liu, B., Xu, H., and Jia, P. 2011. Constrained LDA for grouping product features in opinion mining. In *Advances in knowledge discovery and data mining* (pp. 448-459). Springer Berlin Heidelberg.

[29] Guo, H., Zhu, H., Guo, Z., Zhang, X., and Su, Z. 2009. Product feature categorization with multilevel latent semantic association. In*Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1087-1096). ACM.

[30] Liu, L., Lv, Z., and Wang, H. 2013. Extract product features in Chinese web for opinion mining. *Journal of Software*, *8*(3), 627-632.

[31] Zhai, Z., Liu, B., Xu, H., and Jia, P. 2011. Clustering product features for opinion mining. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 347-354). ACM.

[32] Syed, A. Z., Aslam, M., and Martinez-Enriquez, A. M. 2014. Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text. *Artificial Intelligence Review*, *41*(4), 535-561.

[33] Jo, Y., and Oh, A. H. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 815-824). ACM.