ISSN 2277-3061

# Filtering and Transformation Model for Opinion Summarization

Ms. Ashwini Rao, Dr. Ketan Shah
Assistant Professor IT Dept, MPSTME,Mumbai
ashwini.rao@nmims.edu
Associate Professor, MPSTME, Mumbai
ketan.shah@nmims.edu

## ABSTRACT

The rapid evolution of Micro blogging sites such as Blogs & Twitter facilitate people to post real time messages about their opinions on a variety of topics inclusive of products they use in their daily life. Summarizing opinions of bloggers has several interesting and commercially significant applications like helping the customer to reach purchasing decisions and as a guide for the business activities of companies such as product improvement & market adoption.The paper explores the data characteristics' of Tweets/ Reviews which can be centrepiece of a conversation & provide excellent channel for opinion creation. The short length of the messages and their noisy nature makes it difficult to mine the micro blog data for opinions. Also the infrequent entities such as people, organization, products etc. and user creativity followed by freedom of language hinder the task of Opinion summarization. The paper demonstrates the major role played by Filtering and Transformation techniques in choosing representative words which is the basis for Features extraction in the task of Opinion summarization.

The paper concludes by proposing a framework for pre-processing which emphasises that feature reduction is an important step in Feature based summarization while not compromising on accuracy.

## Keywords

Opinion summarization; Twitter; Social Web; Micro blogging.

ISSN 2277-3061

# I. INTRODUCTION

The web now holds huge volume of data, most of which are opinions of customers about products and services. It is extremely difficult for an individual to manually collect and digest the reviews of his/her interest which would actually help them to reach the right purchasing decisions. In some cases it may also guide the business activity of a company such as improving a product. The vast availability of these opinions has led the researchers in the direction of Opinion Summarization.

It is a sentiment analysis task that helps users to digest the vast availability of opinions in a easy manner. Beyond such summaries, the newer generation of opinion summaries includes structured summaries that provide a well-organized breakdown by aspects/topics. The main aim here is to identify features of a given entity, and summarize the overall sentiment orientation towards each feature. This field of research which goes on to generate effective summaries on every feature of a given entity is called Aspect/Feature based summarization. This category of opinion summarization divides the input texts into features and subtopics. A feature of an entity includes both its components and attributes. For example in a mobile phone the various features/aspects can be its size, audio clarity, battery life etc. Finally it generates summaries of each feature. The aspect based approaches are very popular and have been explored over the last few years [1][2]. This type of summarization is achieved in 3 distinct steps as shown in Fig. 1.

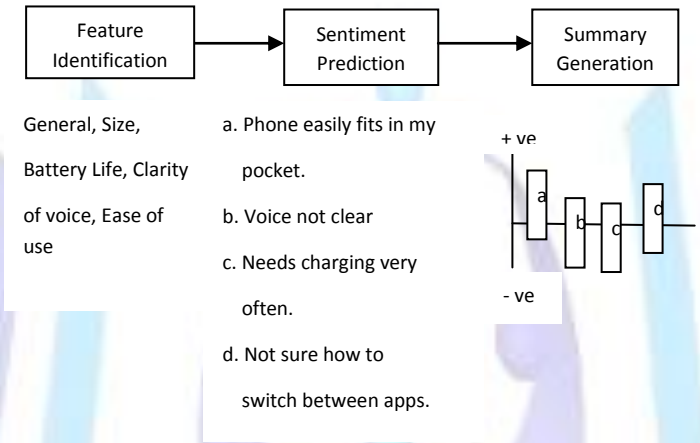


**FIGURE 1. TASKS OF OPINION SUMMARIZATION**

a) Aspect/Feature identification is responsible for identifying salient topics within the text to be summarized. For example if we want to generate an opinion summary about a cell phone some common features are size, battery life, clarity of voice ,ease of use etc.

b) Sentiment prediction which follows feature identification deals with discovering the sentiment orientation (Positive or negative) of the features/aspects found in the previous step.

c) Summary generation is a critical step which makes use of results of feature discovery and sentiment prediction to present the final opinion summaries in a way which is effective and a format which is understandable.

# II. RELATED WORK

Researchers have used multiple approaches towards pre-processing of micro blog data set. The most popular micro blogging site which is used in many sentiment analysis tasks is Twitter. It is popular because of its short length, large volume, diverse and multilingual population as well as its generous nature of Terms of Service. Many researchers have

built corpora of Twitter data. It has been reported that about 500 million tweets are posted every day.

Xinfan Meng et al. [1] have pre-processed their Twitter data by just removing stop words, numbers, URL's and words starting with @. But have concluded that the effectiveness of their entity centric opinion summarization could have been improved using elaborate tokenization methods.

Many research works have been done on summarizing sentiments expressed in Reviews/Blogs. Lun-Wei Ku et al. [2] in his paper on summarization & tracking in news & Blog corpora discuss the role of removing the information which is redundant. Their experimental results showed an improved accuracy if the repeated opinions of same polarity are not dropped. However, some studies show little importance to this cleaning process in their research. For example Yelena Mejova & Padmini Srinivasan et al. [3] in their paper on mining political sentiment in social media concluded that punctuation, special characters need not be removed as well as Stemming need not be performed on the words as tweet specific features such as hash tags and emoticons had to be captured. M. A. Z. Mita, et al. [4] discusses the importance of summarizing noisy unstructured online reviews using extensive pre-processing. They propose a product opinion summarizer which has the first phase to be the pre-processing step. Here the basic cleaning tasks like sentence boundary detection and spell error correction are discussed. An effective method of conflating repetitive punctuation symbols to single occurrence is also experimented.

A sentence merge rule step to handle a situation when user presses unnecessary return keys are proposed by Dey et al. [5] as a unique pre-processing step.Alexander et al. [6] in their paper used Filtering techniques to obtain unigrams which provided a good coverage of the data. They used the process of filtering to remove URL links & twitter special words RT. A set of stop words were also removed using bag of words approach.

Alec Go et al. [7] in their paper on Twitter classification using distant supervision have found a huge reduction of about 45.85% in the size of Feature set. This was made possible by reducing the feature space using the unique properties of Twitter language model. They have demonstrated that few emoticons do not define correct sentiments, so they have stripped of emoticons along with letters which occur repeatedly. Apoorv Agarwal et al. [8] have managed to introduce 2 new resources for pre-processing twitter data. An emoticon & an Acronym dictionary have been prepared. All the emoticons are then replaced by their sentiment polarity & slang words are taken care using Acronym dictionary.

Apart from these there are various other issues like rapidly changing out of dictionary slang, short forms, punctuation error or omissions, phonetic spelling, intentional misspelling & recognition of out of dictionary named entities.

Henriquez et al. [9] in their work proposed an approach which uses n gram based short message text system which was able to correct sentences syntactically from input with a high frequency of misspelled words & internet slang. But the effectiveness of the model had a strong dependency on the quality and size of the dictionary which in turn was not able to handle all possible abbreviations and terms.

Ritter et al. [10] in their work of modelling twitter data, resorted to selecting clusters of spelling variations manually. This approach may not prove to be feasible when we have huge volume of data set which prompts researchers to find effective way of text normalization/pre-processing of casual English.

## III. FEATURES OF TWEETS/BLOGS

As discussed, Twitter is a micro blogging tool which is increasingly popular and allows its users to post messages or Tweets. These tweets are of length 140 characters and are available for immediate download over the internet.

Tweets are extremely useful to marketing as their interaction with public can indicate customer's views about a product far more than traditional media or web pages. The texts published in Twitter have several special features which are typical and issues that rise from them are to be resolved before they can be of any help in Opinion summarization. Some distinctive characteristics of Tweets:

1. Length of Tweets is limited, so users normally use a variety of short & irregular forms of words. This is a unique challenge in sentiment tasks and is called Data sparsity.
2. Linguistic analysis is still more difficult as the user hardly bothers about the grammar that is being used.
3. The usage of creative spelling, punctuations and slang words.
4. The usage of emoticons and emphasizing a word by repeating some letters is also a common trait in tweets.
5. Finally language employed have some special characteristics such as
    a) Symbol RT indicating that the tweets were reposted by other users.
    b) Symbol # (hash) to mark up the topics.
    c) Symbol @ mentioning the users.

Two important sources of opinions other than Tweets are News and Blog articles. The writing of Blogs is more casual when compared to the news articles. Blogs/reviews are written by reviewers in a unstructured way using natural language. Mining through these huge data set is a challenging knowledge engineering task. The highly unstructured characteristic of these online user reviews makes the task of analysing them more difficult.

Major problems are encountered when we try to pre-process these texts. Some related to incorrect punctuations, spelling mistakes, use of slang words, undefined abbreviations etc. Above all the opinions sometimes are expressed in terms of partial phrases rather than complete grammatically correct sentences. So the task of summarizing these noisy online reviews demands extensive pre-processing as well as a technique to first filter out low quality reviews which if not handled may hamper the accuracy of summarization results.

## IV. FEATURE IDENTIFICATION FOR OPINION SUMMARIZATION

Aspect/Feature identification in opinion summarization is the first and a very important step. Currently, a great deal of research has been done on aspect extraction. Identifying aspects and their polarities is very challenging & critical for effectiveness for the task of summarization in opinion mining. This involves extracting fine grained information from opinionated documents. For example, if we want to generate an opinion summary about 'iPod', some of the common aspects are 'battery life', 'sound quality' and 'ease of use'. The purpose of this step is to find these subtopics. These features need to be selected and extracted appropriately as some types of words or phrases do not bear sentiments on their own, but when they appear in some particular contexts, they imply positive or negative opinions. The features thus selected are then stored in feature vectors for further processing. The size of this feature vector would be huge if those features which are unimportant are not filtered.

Various feature extraction approaches like using Part-of-speech (POS) tagger [11] to find nouns & noun phrases, the various language models [12] depend largely on the data set collected. So those tokens which may not reflect any

sentiments have to be removed before they can be used by the above mentioned models. This necessitates the task of filtering and transformations as these features have to be extracted and associated problems had to be solved before sentiment analysis can achieve the next level of accuracy. Any feature of an entity not identified here may lead to summarization results that have low precision and recall. For feature selection the existing solutions can be grouped into 3 categories. One that makes use of language dependency rules [13], next which uses sequence learning algorithms [14] and the last one which makes use of topic models [15].

## V. REQUIREMENT & NEED FOR FRAMEWORK

We collected over 500 reviews of a leading mobile phone from several popular review web sites. We also collected around 350 tweets manually of the same mobile phone.

The collected reviews & tweets were analyzed for their various characteristics as discussed in the previous section. It was observed that in case of tweets, percentage of tokens like # tags, RT tags, & URL's are eminent as they are its main features. Table 1 shows the characteristics of tweets and the percentage of these features is shown in Fig 3.

**TABLE I : FEATURES OF TWEETS**

| **Features** | **No. of tokens** | **% of tokens** |
|---|---|---|
| Others | 10915 | 70% |
| # tags | 1570 | 10% |
| URL's | 1099 | 7% |
| Stopwords | 1692 | 11% |
| RT tags | 139 | 1% |
| Special characters & digits | 126 | 1% |
| **Total** | **15541** | 100% |

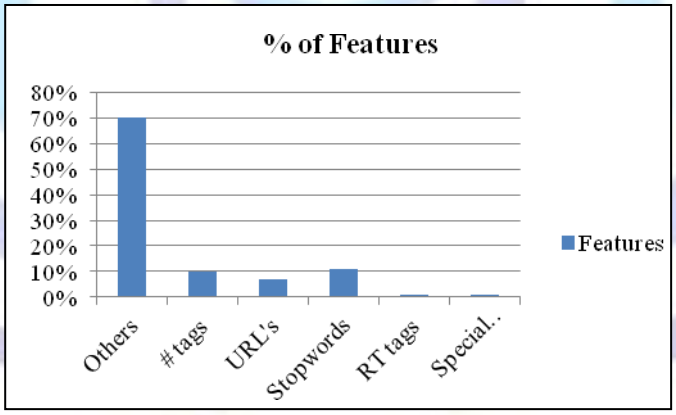


**FIGURE 3. CHARACTERISTIC FEATURES OF TWEETS**

The characteristics of Reviews are shown in Table 2 and the percentage of its features in actual data set is depicted in Fig. 4. We observed that reviews have less number of URL's as well as the usage of slang words and abbreviations is also less when compared with tweets. The reason being that they are usually written by experts and do not have any restriction when it comes to their length. But the number of stop words is comparatively higher than tweets again because of the same reason mentioned previously.

ISSN 2277-3061

**TABLE 2**

**FEATURES OF REVIEWS**

| Features | No.of Tokens | % of Tokens |
|---|---|---|
| Other Tokens | 31642 | 55% |
| No. of URL's | 171 | 0% |
| No. of Stopwords | 25539 | 44% |
| No. of special characters & digits | 444 | 1% |
| **Total** | **57796** | 100% |

So, around 30% of the tokens in tweets and 45% of them in reviews do not carry any sentiments. But out of the 30% & 45% tokens, there are certain stop words like 'not', 'and', 'but', etc. which are very important when it comes to classification of sentiments. So all of these words cannot be considered as candidates for filtering.

This percentage of unwanted tokens will surely rise when data sets becomes huge. This further result in the increase of feature vector size and thereby making the task of feature selection more difficult. We performed experiments on these noisy data sets by applying some filtering and transformation techniques as discussed in the next section.

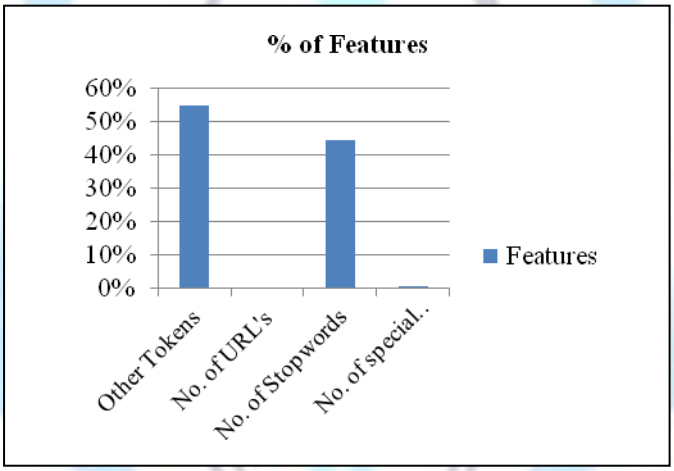


**FIGURE 4. CHARACTERISTIC FEATURES OF REVIEWS**

## VI. EXPERIMENTS DONE & RESULTS

As discussed in the previous section, because of the unstructured characteristics of the micro blog data set collected, we conducted experiments on them by applying various filtering techniques and then some transformations. We evaluated the percentage reduction in feature space after the various filtering tasks were applied on the twitter & the review data set.

As discussed previously, because of the importance of some stop words, all of them are not filtered. So a stop word dictionary is constructed manually based on the characteristic of data collected rather than using a freely available general stop word list. As depicted in Fig 5. & Fig 6.we observed around 23% reductions in feature space of Twitter data set and about 17% reduction in Reviews data set respectively even though the stop words are not totally removed.

ISSN 2277-3061

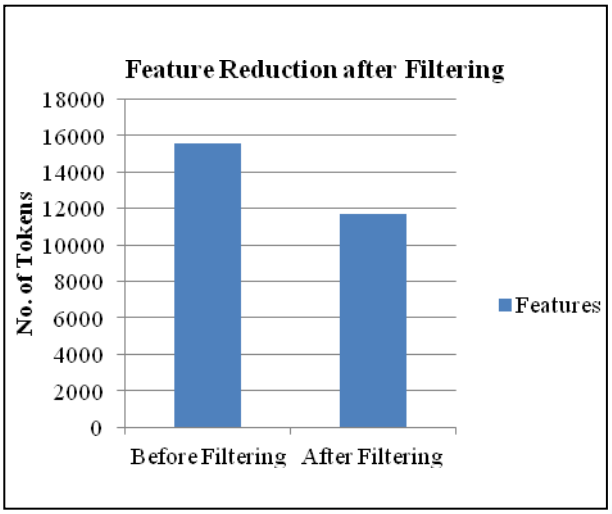


**FIGURE 5 . FEATURE REDUCTION IN TWEETS**

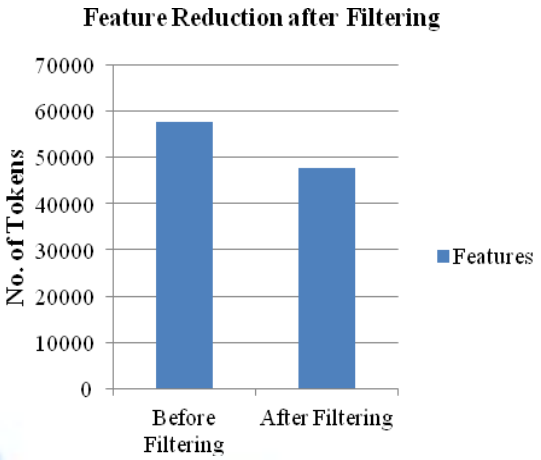


**FIGURE 6. FEATURE REDUCTION IN REVIEWS**

As it is quite evident from the results obtained, filtering plays a vital role in the reduction of feature vector space and role of transformation becomes clear in the classification step which would finally give a good accuracy to the task of Feature based summarization. So a framework model is proposed in the following section.

## VII. ANALYSIS & FRAMEWORK MODEL

As per the experiments conducted and results obtained in the previous section, we propose a Framework model as shown in Fig. 7 for pre processing. This model would be used for the task which precedes feature extraction. The main aim of this step is to structure the text into an appropriate form which can be further summarized with a better accuracy. The following are the various Transformations and Filtering carried out.

1. Fully capitalized words are converted to their lower case counterparts.
2. Retweets (RT), username@x, and URL links are removed as they express no sentiments.
3. Hash tags(#) have also been removed.
4. Repeated letters have been reduced (e.g. cooool by cool, niiice by niice). These are not corrected to their actual word in order to have a effect of these emphasized words on the accuracy of opinion summarization.
5. English contraction words like don't, can't etc. are converted into do not and cannot respectively.
6. Boundary of sentences is detected and a white space is inserted between the last word and punctuation symbol. This is done to demarcate the sentences which can be either interrogative sentences (those ending with ?) or imperative sentences (those that start with verbs). Such interrogative sentences are then removed as they are more likely to be neutral.

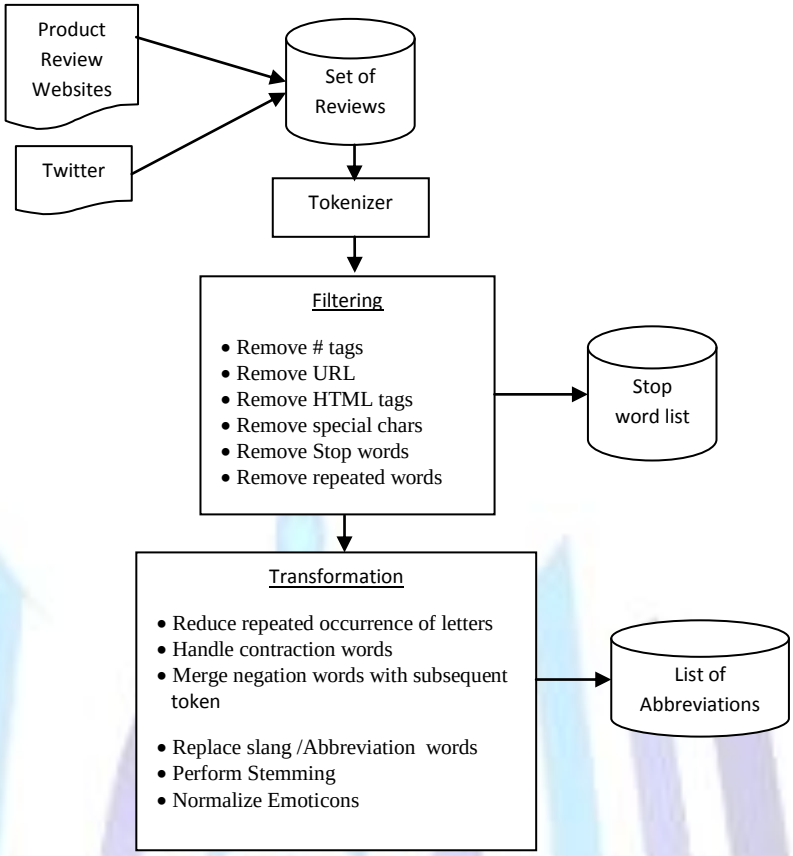


**FIGURE 7. PROPOSED FRAME WORK MODEL**

1. Some of the unnecessary characters like $,%,{},(),*+,<,=,> are also removed along with digits/numbers which carry no sentiments.
2. Whenever the token 'not' is encountered it is merged with the subsequent token. This boosts the performance as handling negation casually may reverse the polarity of the sentiments to be summarized altogether.
3. All the HTML tags are removed.
4. The various tokens that demonstrate laughter like ha,ha..he,he..are replaced by the token LAUGH.
5. The various emoticons like :) , ;), :-) which express positive feeling are normalized to token HAPPY and all those which indicate negative feeling like :(, ;( , :-( are normalized to token SAD. Also some emoticon

   tokens are replaced by emoticon words using an emoticon dictionary obtained from Wikipedia.
6. In tweets many punctuation marks like!(exclamation), " & " may indicate some strong feelings. So these are normalized as SPECIAL along with their frequency of occurrence.
7. Abbreviations or slang words like lol (laughing out loud), gr8 (great),bff (best friend forever) etc.. are replaced by their original form using a list of slang words derived from social media.
8. Stemming is performed to replace non stemmed words such as lover, loved, loving to love. This is done using Porter stemmer. If not these non stemmed

   words will create problems when search needs to be done using some dictionaries in later stage of summarization tasks.

## VIII.Observation & Conclusion

We analyzed that even though summarizing opinions of Tweeters and Bloggers has many interesting and significant applications, they may not give accurate results if intensive preprocessing is not done on them. We have found out that the importance of the filtering tasks is much more crucial in twitter data than in reviews. This is mainly because the reviews are written by professional reviewers who hardly make use of a writing which is highly unstructured. So a general frame work model for preprocessing cannot be proposed, as the tasks have to be carried out based on the characteristics of the data set. We have also observed the sentiments in tweets are a topic which is worth further investigation.

As the future work, we plan to perform preprocessing on larger and more varied micro blog data sets. We would also like to explore the effect of the proposed model in the feature generation and its role in achieving better accuracy in Feature based summarization.

## REFERENCES


[1] F. W. Xinfan Mengz, Xiaohua Liuy, Ming Zhouy, Sujian Liz, Houfeng Wangz, "Entity-Centric Topic-Oriented Opinion Summarization in Twitter," In KDD, 2012.

[2] Y.-T. L. Lun-Wei Ku, Hsin-Hsi Chen, "Opinion Extraction, Summarization and Tracking in News and Blog Corpora," In Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006.

[3] P. S. Yelena Mejova, Bob Boynton, "GOP Primary Season on Twitter: "Popular" Political Sentiment in Social Media," In WSDM, 2013, pp. 517-526.

[4] M. A. Z. Mita K. Dalal, "Semisupervised Learning Based Opinion Summarization and Classification for Online Product Reviews," Applied Computational Intelligence and Soft Computing, p. 8, 2013.

[5] L.Dey and S. M.Haque, "Opinion mining from noisy text data," International Journal on Document Analysis and Recognition,vol. 12, no. 3, pp. 205–226, 2009.

[6] P. P. Alexander Pak, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," In International Language Resources and Evaluation, 2010, pp. 1320-1326.

[7] R. B. Alec Go, Lei Huang, "Twitter Sentiment Classification using Distant Supervision," Stanford University.

[8] B. X. Apoorv Agarwal , Ilia Vovsha ,Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data," In LSM '11 Proceedings of the Workshop on Languages in Social Media, 2011, pp. 30-38.

[9] Henriquez CA, Hernandez ," A ngram-based statistical machine translation approach for text normalization On chat-speak style communications ," In Proceedings of CAW2.0, Madrid, Spain, August 2009, 1–5.

[10] Ritter A, Cherry C, Dolan B,"Unsupervised modeling of Twitter Conversations," In Proceedings of HLT- NAACL 2010, Los Angeles, California, June 2010, 172–180.

[11] Hu, M. and B. Liu. Mining opinion features in customer reviews. In Proceedings of National Conference on Artificial Intelligence (AAAI-2004), 2004a.

[12] Mei, Q., X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In Proceedings of International Conference on World Wide Web (WWW-2007), 2007.

[13] Qiu, G., B. Liu, J. Bu, and C. Chen. Opinion word expansion and target extraction through double propagation. Computational Linguistics, 2011.

[14] Jin, W. and H. Ho. A novel lexicalized HMM-based learning framework for web opinion mining. In Proceedings of International Conference on Machine Learning (ICML-2009), 2009a.

[15] Titov, I. and R. McDonald. Modeling online reviews with multi-grain top-ic models. In Proceedings of International Conference on World Wide Web (WWW-2008), 2008a.

[16] Kukich K ," Techniques for automatically correcting words in text," ACM Computing Surveys, 24(4), 377–439.

[17] Clark E, Roberts T, Araki K," Towards a Pre- processing System for Casual English Annotated with Linguistic and Cultural Information," In Proceedings of Computational Intelligence 2010, Hawaii, August 2010.

[18] B. Liu, "Sentiment Analysis and Subjectivity," Handbook of Natural Language Processing, Second ed., 2010.

[19] Godbole, N. Srinivasaiah, M. Skiena,"Large-scale sentiment analysis for news and blogs," In Proceedings of the International Conference in Weblogs and Social Media, 2007.

[20] Gamon, M,. "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," In Proceedings of the International Conference on Computational Linguistics (COLING), 2004.

[21] Twitter API, " https://dev.twitter.com/docs/streaming-apis ".

[22] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011 Workshop on Languages in Social Media,2011 , pp. 30–38.

[23] A. Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1–15. Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems.