# Brief Survey on Opinion Mining Techniques

Swapnil B. Mulay
PG Student
Affiliated to Savitribai
Phule Pune University, Pune
Department of Computer Engineering
DY Patil College of Engineering,
Ambi, Pune, India

Dhanashree Kulkarni
Professor
Affiliated to Savitribai
Phule Pune University, Pune
Department of Computer Engineering
DY Patil College of Engineering,
Ambi, Pune, India

## ABSTRACT

The term review mining and opinion mining is gaining its interest with the expanded use of web. Whatever user buys or read or sees on the web, he likes to write his opinion about it or read others reviews to get needed information. Such websites also enforce users to write their reviews about the topic or product. It is common to express the opinion about the products by the user on websites of the products. It is of interest of both the company and the users as these are source of getting feedback and suggestions. Various schemes are proposed in literature to extract useful information from these reviews. Most important step in opinion mining is extraction of opinion and target words. Target depicts about what opinion is given and opinion words are the words used to express the opinion about target. Different techniques are developed so far, for efficient and accurate extraction of target and opinion words. Due to importance of opinions of the product by other users in decision making of user, number of fake reviews is increased. We proposed system to detect and remove fake comments before extraction process which classifies fakes comments and extracts opinion, target words only from genuine comments.

## General Terms

Opinion and target word extraction, fake review detection.

## Keywords

Opinion mining, opinion targets extraction, opinion words extraction.

## 1. INTRODUCTION

Nowadays it is common to express the opinion about the products by the user on websites of the products. It is very useful to the company of the products and also to other users who are thinking to buy that product. From company's perspective, online reviews are easy source to get feedback and suggestion about products for improvement. Due to its importance in the industry many researchers worked on it.

Generally review can't reflect the overall sentiment of user about product because each review is combined of one or many dimensions and each dimension's description may hold distinct sentiment. For example take one review about Camera "Great quality of pictures but camera has battery problem. Display on camera is average", in this review we can't conclude overall sentiment of user about camera. Review has three dimensions, Quality of pictures, Battery, Display and sentiment of each is different. To analyze these reviews it is needed to extract words about which user is writing and these words are called as Opinion target (Quality of pictures, Battery, Display). There is need to extract words which describe opinion of users about the opinion target called as Opinion word (Great, Problem, Average).

Most of the methods for extraction of opinion and target words proposed in the literature depend on the collective extraction strategy. Collective strategy is based on the assumption that opinion words are frequently comes with target words and there is strong association among them. For example consider a product mobile phone and reviews on it. "Big" and "colorful" words are used to describe the "screen". Therefore if we know review contains the "Big" as opinion word then there is high possibility "screen" is the opinion target. Extraction process is performed alternatively between target and opinion word until there is nothing to extract.

Traditional methods for extraction are suffered from several limitations. One of the most used methods for extraction was nearest neighbor rules in which nearest adjective/verb to a noun/noun phrase in a limited window is considered as its modifier. Review may hold long-spam modified relations and various opinion expressions therefore this method cannot get precise results. After this method many syntactic patterns were designed which the opinion relations among words are decided according to their dependency relations in the parsing tree. No one can guarantee grammatical correctness of the online given by users and short forms, smileys are used largely in the reviews. Syntactic patterns will give good results against formal text but cannot handle grammatically wrong reviews. Collective extraction methods are based on bootstrapping framework in which error is propagated to the last result. In bootstrapping method consideration is, if one word is target word then word with that word is opinion word. In case if prior knowledge is wrong then this error is propagated to all next iterations. To overcome such results method[6] designed for extraction of opinion and target word which is based on partially supervised word alignment model and graph co-ranking. From experimental results it is clear that proposed method [6] outperforms state-of-the-art methods.

Extracted target and opinion words are then used for sentiment analysis. Based on the result of the sentiment analysis many decisions are taken to improve the product and business. There is indirect but considerable impact of reviews for decision making in business and potential users also assess the quality of the product by reading reviews. Considering importance of the review, adversaries attempt to reduce the reputation of the product by fake reviews. If there are present fake reviews then it will surly affect the sentiment analysis negatively. For accurate sentiment analysis there is need of only genuine reviews. To overcome this issue paper proposed a fake review detection technique to detect and remove fake comments before extraction process which classifies fakes comments and extracts opinion, target words only from genuine comments.

## 2. LITERATURE SURVEY

[1] Web opinion mining is a topic that gaining interest for businesses that run online. Different websites are available for various purposes. Users of such websites are pushed to write their opinion and reviews about the content of the website or main prospect for what the website is built. Different reviewers have contrasting opinion about the respective topic. Before making a buying decision, users tend to depends on reviews about the product. This paper considered problem of review mining and categorized it into subtasks, such as; determining features of the product and opinion about product features, identifying dimensionality in opinions, based on opinion strength ranking them accordingly. Paper proposed unsupervised extraction system, called OPINE that work out above discussed problems. System works in the context of specific review sentences and extracted product features. Technique determines opinions of the customers and dimensionality of opinions with utmost exactness.

[2] This paper addressed the issue of untruthful reviews on movies on IMDB website and proposed a system to detect fake reviews on the websites. Proposed method used J48 classifier to classify comment as fake or genuine. Beyond detection of fake reviews method also tries to find brand spamming, in which particular brand is promoted in fake reviews.

[3] Opinion analysis with user sentiment analysis provides much valuable information for opinion mining. Sentiment analysis is the analysis of user's expression and attitude towards the event. Paper uses the supervised approach for opinion target extraction, when the available number of labeled data is low then model is not trained properly and gives inefficient results. To address this problem, paper considered sentiment analysis for knowledge extraction and proposed supervised learning domain adaptation method works in case of labeled data lacking. The introduced technique that first determines sentiment words within domain and source and then combines them. Then to derive relation amongst topic and sentiment, an algorithm is proposed known as Relational Adaptive BootstraPping (RAP). Experiments showed that technique used extraction of opinion words and opinion targets with high precision.

[4] Product reviews contain users experience description about the product. This helps Business organizations to form future policies regarding the product and its marketing. And also customers intend to buy the product can make better decision whether to go for it or not. Reading through thousands of reviews and extracting appropriate information to make a buying decision is not possible. For example, consider a book as an object then its positive opinion would be great, interesting etc. and its negative opinion would be boring, bad etc. This requires 1) determining all objects features and opinions and 2) determining dimensionality in sentiment expression. Paper considered draw up these two tasks as joint structure tagging problem and a framework based on Conditional Random fields was proposed. Here Structure-Aware means modeling the relationship among output labels. Proposed framework can exploit the relationship between objects and positive negative opinions and is able to integrate naturally linguistic information into model representation. Results showed that proposed framework is effective in review mining.

[6] Opinion target extraction which is an important and main task of opinion mining contains task of extracting items on which the opinions are based. Opinion targets are the object of interest about which people expressing their opinions. For example customer is writing about screen of the mobile in wideness, here mobile screen is opinion target. Opinion word is the view of users about particular object. For example screen resolution is awesome, here awesome is opinion word. Opinion words and target are closely related to each other and have strong association. To depict the relation between these two this paper proposed WAM (Word Alignment Model). Also for extraction paper introduced partially supervised WAM model in case of quality degradation of alignment. This method is able to identify relationship among the words in sentences. In the case of syntax pattern parsing of sentences is needed, in proposed technique there is no requirement of such need to parse the informal sentences. Proposed model shows the relationship between words. It gives accurate results by identifying opinion target and word relation.

[7] Double Propagation technique for feature mining gives less recall and less precision for large corpora. To improve recall part whole and no patterns are introduced to increase the recall. Importance of the feature is evaluated using feature relevance and frequency. Features are ranked using importance of the feature using HITS ranking algorithm. Experiments on the Cars, mattress and Phone dataset shows increase in the Precision and recall in proposed approach than Double propagation method.

[9] Paper proposes number of feature mining technique with aim to generate feature based summary of the number of reviews posted by the customers. Feature extraction done by various methods frequent features, compactness pruning, p-support pruning, and Infrequent feature identification. Each feature extraction technique was applied on dataset of 5 different products and precision and recall was evaluated. Average recall and average precision was 80% and 70 % respectively

[11] Extraction of opinion words and target words from comments on the news article was the aim of this paper. In this context, many times comments are irregular and informal and sometimes opinion target is implicit. To solve this problem centering theory was applied. Accuracy of target extraction is used for evaluating the effectiveness of the proposed method. Accuracy is ratio of number of sentences with correct extraction and total number of extraction. Proposed method gives highest accuracy 43.20 than all other existing methods (baseline 1, 2 and FC only (Focused only on Concepts)) at that time.

[18] In opinion mining, target word mining is an interesting aspect and also a challenging task. This paper proposed an approach, a partially supervised word alignment model for opinion target extraction (PSWAM). This technique finds opinion relations within sentences and tried to predict association between the words. Then confidence of each candidate is estimated and candidate with greatest confidence is considered as opinion target, this is achieved using a graph-based algorithm. Proposed technique is effective as it easily avoids parsing errors and also efficiently deals with informal sentences within online reviews. Three different size datasets are used for experiments showed the method performed better.

**Table 1. Study of literature along with advantages/disadvantages.**

| **Paper name and authors** | **Description** | **Result / Advantage / Disadvantages** |
|---|---|---|
| Extracting product features and opinions from reviews<br>A.M. Popescu and O. Etzioni 2005[1] | Web opinion mining is a topic that gaining interest for businesses that run online. Different websites are available for various purposes. Customers write reviews about what they are feeling about the particular product they have bought online, or may write about any topic, any person or any aspect. | System works in the context of specific review sentences and extracted product features. Technique determines opinions of the customers and dimensionality of opinions with utmost exactness |
| Fake Review and Brand Spam Detection using J48 Classifier<br>Sushant Kokateand Bharat Tidke<br>2015 [2] | This paper addressed the issue of untruthful reviews on movies on IMDB website and proposed a system to detect fake reviews on the websites. Proposed method used J48 classifier to classify comment as fake or genuine. | Beyond detection of fake reviews method also tries to find brand spamming, in which particular brand is promoted in fake reviews. |
| Cross-domain co-extraction of sentiment and topic lexicons<br>F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu 2012[3] | It considers sentiment analysis for knowledge extraction and proposed supervised learning domain adaptation method. And introduced technique that first determines sentiment words within domain and source and then combines them. Relational Adaptive BootstraPping (RAP) algorithm is proposed to derive relation amongst topic and sentiment. | Experiments showed that technique used extraction of opinion words and opinion targets with high precision. |
| Structure-aware review mining and summarization." F. Li, C. Han, M. Huang, X. Zhu, Y. Xia, S. Zhang, and H. Yu 2010 [4] | Paper considered draw up target and opinion word extraction task as joint structure tagging problem and a framework based on Conditional Random fields was proposed. Here Structure-Aware means modeling the relationship among output labels | Proposed framework can exploit the relationship between objects and positive negative opinions and is able to integrate naturally linguistic information into model representation. Results showed that proposed framework is effective in review mining. |
| Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model Kang Liu, Liheng Xu, and Jun Zhao 2015 [6] | Opinion words and target are closely related to each other and have strong association. More opinion words can be derived using extracted opinion targets. To depict this relation WAM (Word Alignment Model) is proposed. Also for extraction paper introduced partially supervised WAM model in case of quality degradation of alignment. | Proposed model shows the relationship between words. It gives accurate results by identifying opinion target and word relation |
| Extracting and Ranking Product Features in Opinion Documents<br>L. Zhang, B. Liu, S. H. Lim<br>2010 [7] | Double Propagation technique for feature mining gives less recall and less precision for large corpora. To improve recall part whole and no patterns are introduced to increase the recall. Importance of the feature is evaluated using feature relevance and frequency. Features are ranked using importance of the feature using HITS ranking algorithm. | Experimental results on the Cars , mattress and Phone dataset shows increase in the Precision and recall in proposed approach than Double propagation method. |
| Mining Opinion Features in Customer Reviews<br>Minqing Hu and Bing Liu2004 [9] | Paper proposes number of feature mining technique with aim to generate feature based summary of the number of reviews posted by the customers. Feature extraction done by various methods - frequent features, compactness pruning, p-support pruning and Infrequent feature identification. | Each feature extraction technique was applied on dataset of 5 different products and precision and recall was evaluated. Average recall and average precision was 80% and 70 % respectively. |
| Opinion Target Extraction in Chinese News Comments<br>Tengfei Ma and Xiaojun Wan 2010 [11] | Extraction of opinion words and target words from comments on the news article was the aim of this paper. In this context, many times comments are irregular and informal and sometimes opinion target is implicit. To solve this | Accuracy of target extraction is used for evaluating the effectiveness of the proposed method. Accuracy is ratio of number of sentences with correct extraction and total number of extraction. Proposed method gives |

| | | |
|---|---|---|
| | problem centering theory was applied. | highest accuracy 43.20 than all other existing methods (baseline 1 , 2 and FC only (Focused only on Concepts)) at that time |
| Opinion target extraction using partially-supervised word alignment model,<br>K. Liu, H. L. Xu, Y. Liu, and J. Zhao, 2013 [18] | This paper proposed an approach, a partially supervised word alignment model for opinion target extraction. | Proposed technique is effective as it easily avoids parsing errors and also efficiently deals with informal sentences within online reviews. Three different size datasets are used for experiments showed that the method performed better. |

## 3. PROPOSED SYSTEM

### 3.1 Training data

Training data consist of reviews of the product with labels i.e. whether review is Fake or genuine. This training data is collected from internet and domain expert label then by analyzing them manually.

### 3.2 Testing data

Testing data is the reviews without labels and need their label for further extraction procedure. All reviews of product or item whose sentiment analysis is done are considered as Training data.

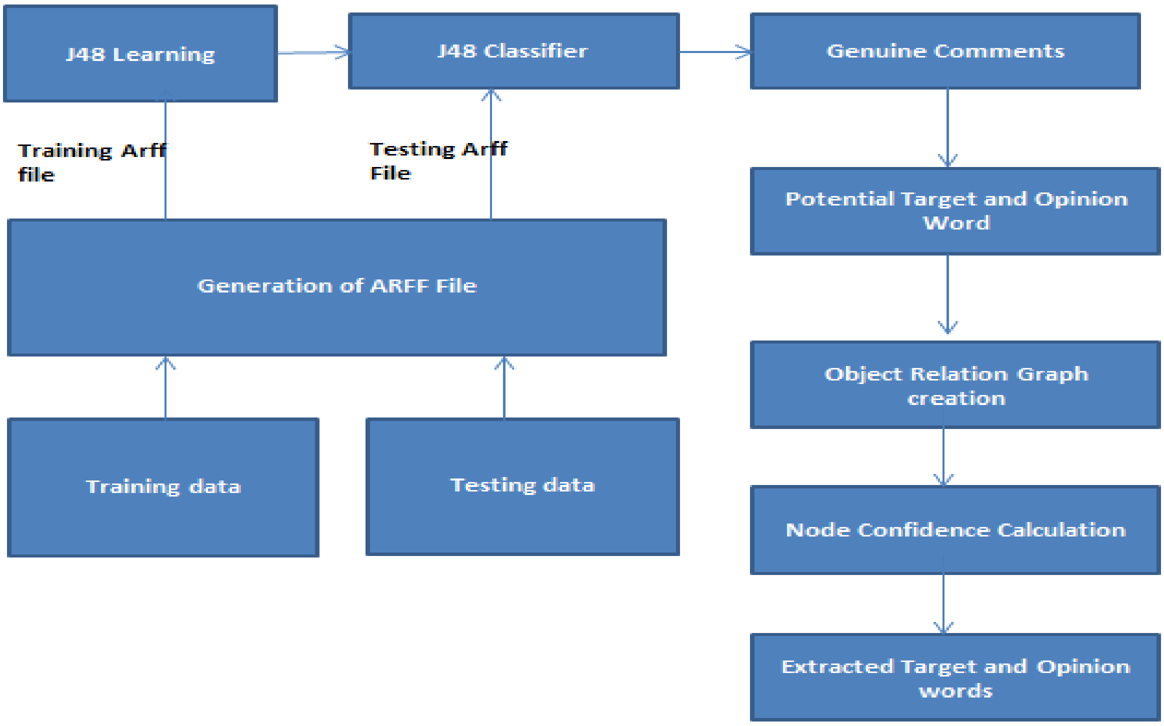

**Fig. 1: Proposed System**

### 33 Generation of Attribute Relation File Format (ARFF) file from testing or training data

Each review is converted into instance in the ARFF file. Instance in the ARFF is vector of attribute values. For example Patient is relation and age, weight, height, gender and class are the attributes. Class is attribute which has values positive or negative. 50, 65, 165, male, positive is example of the instance where 50, 65,165 and male are values of attributes and positive is class value. Each review is converted into instance. Class values are kept blank for testing reviews. Following attribute values are extracted from reviews to form instance in the arff file.

### 34 J48 Training

J48 is weka implementation of C4.5 decision tree classifier. Arff file generated from training files is called as training arff file. Training arff file given to J48 classifier as input. J48 classifier builds decision tree as per data in the training arff file and used as classifier for further process.

### 35 Testing

In this step decision tree built in the above step and arff file of testing reviews are used and label of the testreviews are predicted. In this step fake and genuine reviews are partitioned as per predicted label. Only genuine comments are used for further extraction process.

## 36 Extraction of Potential Target and Opinion words

This is the first step in extraction of target and opinion words from genuine reviews. Existing system assumes that nouns/noun phrases in sentences are opinion target candidates, and all adjectives/verbs are regarded as potential opinion words. Part of Speech tagger is used for extracting noun, adjectives from sentences. This step gives list of potential target and opinion words.

## 37 Object relation graph creation

OPG is bipartite undirected graph G = (V, E, W) where V = Vt U Vo, V is the set of vertices. V is union of set of opinion word vertices Vo and set of Target word vertices Vt. E is the set of edges between the Opinion word vertex and Opinion target vertex. All potential target and opinion words from previous steps are represents vertices in the G. Edge between two vertices represents that there is opinion relation between corresponding vertices. To spot the relationship between target and opinion partially supervised word alignment model is used. Once the relation between words is identified edge is added in between respective vertices. Weight of the edge denotes degree of association in two vertices. Alignment probabilities between a potential target Wt and Potential opinion word Wo is calculated using,

P(Wt|Wo) = Count(Wt,Wo)/Count(Wo) and Opinion association between Wt and Wo calculated as follows:
OA(Wt,Wo)=(a * P(Wt|Wo)) + 1 / ((1-a) P(Wo|Wt)) where a is harmonic factor used to combine alignment probabilities.

## 38 Node Score Calculation

In this step, potential score of each opinion word andtarget word is calculated using following formula

Conf (Wt,K+1) = (1-u) X OA(Wt,Wo) X Conf(Wo,k) + u X It.
Conf (Wo,K+1) = (1-u) X OA(Wt,Wo) X Conf(Wt,k) + u X Io.

where Conf(Wt,K+1) is Confidence of Potential target word Wt as target word at K+1 round,Conf(Wo,K+1) is Confidence of Potential opinion word Wt as opinion word at K+1 round,Conf(Wo,k) is Confidence of Potential target word Wt as target word at Kth round, Conf(Wt,k) is Confidence of Potential opinion word Wt as opinion word at Kth round, It and Io denote prior knowledge of candidates being opinion targets and opinion words u is impact of the prior knowledge u ϵ (0,1).Opinion and target vertices in the graph whose score is above threshold are extracted as target and potential words.

## 4. CONCLUSION

In this paper, first it has elaborated various opinion and review mining and extraction techniques. With increasing number of online businesses, competition amongst the same is heightened. There is possibility of opinion spamming i.e. fake review writing or promoting some other products or posting some unrelated topic. Untruthful reviews can be made to defame or lower the goodwill of the particular product. For accurate sentiment analysis there is need of only genuine reviews. To overcome this paper proposed fake review detection to detect and remove fake comments before extraction process which classifies fake comments and extracts opinion, target words only from genuine comments.

As this paper considers the most frequently appearing pattern of fake comments and its attribute consideration for detecting the fake comments. Further we would like to enhance the selection process of genuine reviews by considering more attributes which will contribute for detecting more accurate and precise genuine comments. And which will cover most of the fake review patterns. And which will further improve the overall opinion mining process and this extracted opinion and target words can be used efficiently for sentiment analysis process.

## 5. ACKNOWLEDGMENTS

It is been rightly said that we are built on the shoulder of others. For everything I achieved, the credit goes to all those who had really helped me to complete this survey successfully. I am extremely thankful to my Project Guide "**Prof. Dhanashree Kulkarni**" for guidance and review of this paper work. I would also like to thank all the faculty members of "**D.Y. Patil College of Engineering**", and my friends who helped me for this survey.

## 6. REFERENCES

[1] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in Proc. Conf. Human Lang. Technol. Empirical Methods Natural Lang. Process., Vancouver, BC, Canada, 2005, pp. 339–346.

[2] Sushant Kokate, Bharat Tidke, "Fake Review and Brand Spam Detection using J48 Classifier," in (IJCSIT) International Journal of Computer Science and Information Technologies, Vol.6(4),2015. 3523-3526.

[3] F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu, "Cross-domain co-extraction of sentiment and topic lexicons," in Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Jeju, Korea, 2012, pp. 410–419

[4] F. Li, C. Han, M. Huang, X. Zhu, Y. Xia, S. Zhang, and H. Yu, "Structure-aware review mining and summarization." in Proc. 23th Int. Conf. Comput. Linguistics, Beijing, China, 2010, pp. 653–661

[5] A. Mukherjee and B. Liu, "Modeling review comments," in Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Jeju, Korea, Jul. 2012, pp. 320–329.

[6] Kang Liu, Liheng Xu, and Jun Zhao, "Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model." IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 3, MARCH 2015

[7] L. Zhang, B. Liu, S. H. Lim, and E. O'Brien-Strain, "Extracting and ranking product features in opinion documents," in Proc. 23th Int Conf. Comput. Linguistics, Beijing, China, 2010, pp. 1462–1470.

[8] K. Liu, L. Xu, and J. Zhao, "Opinion target extraction using word-based translation model," in Proc. Joint Conf. Empirical Methods Natural Language Process Computer Natural Lang. Learn., Jeju, Korea, Jul 2012, pp. 1346–1356.

[9] M. Hu and B. Liu, "Mining opinion features in customer reviews," in Proc. 19th Nat ConfArtifIntell., San Jose, CA, USA, 2004, pp. 755–760.

[10] G. Qiu, L. Bing, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," Computer Linguistics, vol. 37, no. 1, pp. 9–27, 2011.

[11] T. Ma and X. Wan, “Opinion target extraction in Chinese news comments.” in Proc. 23th Int. Conf. Compute. Linguistics, Beijing, China, 2010, pp.782–790.

[12] Q. Zhang, Y. Wu, T. Li, M. Ogihara, J. Johnson, and X. Huang, “Mining product reviews based on shallow dependency parsing,” in Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, Boston, MA, USA, 2009, pp. 726–727.

[13] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts, “Learning Word Vectors for Sentiment Analysis,” Proc. 49th Ann. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies, pp. 142-150, 2011.

[14] W. X. Zhao, J. Jiang, H. Yan, and X. Li, “Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid,” in Proc. Conf. Empirical Methods Natural Lang. Process., Cambridge, MA, USA, 2010, pp. 56–65.

[15] A. Yessenalina and C. Cardie, “Compositional Matrix-Space Models for Sentiment Analysis,” Proc. Conf. Empirical Methods in Natural Language Processing, pp. 172-182, 2011.

[16] David M. Blei, Andrew Y. Ng, Michael I. Jordan, ”Latent Dirichlet Allocation,” Journal of Machine Learning Research 3, 2003.

[17] Z. Liu, X. Chen, and M. Sun, “A simple word trigger method for social tag suggestion,” in Proc. Conf. Empirical Methods Natural Lang. Process., Edinburgh, U.K., 2011, pp. 1577–1588.

[18] K. Liu, H. L. Xu, Y. Liu, and J. Zhao, “Opinion target extraction using partially-supervised word alignment model,” in Proc. 23$^{rd}$ Int. Joint Conf. Artif. Intell., Beijing, China, 2013, pp. 2134–2140.

[19] Z. Hai, K. Chang, Q. Song, and J.-J. Kim, “A Statistical Nlp Approach for Feature and Sentiment Identification from Chinese Reviews,” Proc. CIPS-SIGHAN Joint Conf. Chinese Language Processing, pp. 105-112, 2010.

[20] Z. Hai, K. Chang, J.-J. Kim, and C. C. Yang, “Identifying features in opinion mining via intrinsic and extrinsic domain relevance,” IEEE Trans. Knowledge Data Eng., vol. 26, no. 3, p. 623–634, 2014.