ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 2, February 2018

A Review of Opinion Extraction and Analysis

[1] S. C. Nandedkar, [2] J. B. Patil, [3] P. A. Joshi
[1][3] Asst. Prof., DIEMS Aurangabad, [2] Principal, R. C. Patel COE, Shirpur.

***Abstract: The* internet is becoming user centric and people preferring to exchange opinions through online social means such as – discussion forums, blogs and micro-blogs. This user opinion base is a valuable resource from buyers, as well as sellers' perspective. It is helpful for buyers to choose a good product and helpful for sellers to improve their product. This is the task known as opinion mining and analysis. This paper represents a simple approach for understanding customers review using machine learning technique. It also illustrates the steps: data gathering, preprocessing, POS tagging, dependence analysis, opinion and target word finding. It also gives the technique for measuring the performance of the above stated system. Finally, it creates the knowledge base to fulfill the said intention.**

***Keywords:* - Data mining, opinion mining, sentiment analysis, unstructured corpus.**

I. INTRODUCTION

The Internet is becoming increasingly user-centric because of emerging communication platform Web 2.0. People prefer exchanging opinions through online social means such as – discussion forums, blogs and microblogs (e.g. facebook, Twitter, customer reviews etc.). Along with such trends, a huge amount of user-generated content is present on the Internet [1]. It consists of rich opinion and sentiment information. We now have a huge volume of this opinion data. A proper understanding and analysis of this opinion and sentiment information have become increasingly important and key influencer of our behavior. For service / product providers, it is important to know customers' feedback for quality improvement. For customer / user of the service, it is important to know others' feedback to find out which is the best product / service [2]. The process is named as Opinion Mining (OM) or Sentiment Analysis. As given in [3], there is a subtle difference in these two terms.

Opinion Mining: It can be defined as sub-discipline of computational linguistics that focuses on extracting people's opinion from the Web. Given a piece of text, opinion mining system analyses

- Which part is opinion expression?
- Who wrote the opinion?
- What is being commented?

Sentiment Analysis: It is about determining the subjectivity, polarity (positive / negative) and polarity strength (weakly positive, mildly positive, strongly positive) of the piece of text.

- What is the opinion of the writer?

Now, both the customers as well as merchants of the product / service need a precise opinion from the customer's opinion. Two major players are the opinion word and its corresponding target word is to be focused on mining purpose. However, it is necessary to have prior knowledge of opinion word lexicons [2]. Opinion target is defined as the object about which user expresses their views. It is also called as features [4]. For example, features for a mobile handset can be screen, battery, resolution, processing speed, etc. Feature extraction means finding out customer comments related to these features. For achieving the above mentioned goal we use Part of Speech (POS) analysis. In this case noun, adjective, verbs and adverbs are the basic form of sensing opinion. As the opinion words usually co-occur with the opinion target, a collective extraction strategy is adapted here. It follows supervised learning approach, i.e. classification with fixed target classes [2, 4, 5, 6].

II. LITERATURE REVIEW

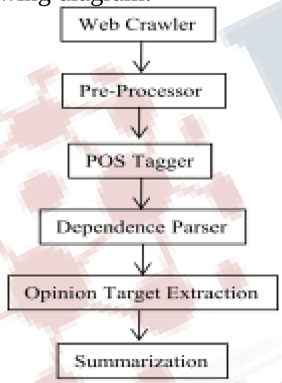
Kang Liu et al. [4] analyzed the relationship between opinion targets and opinion words. This paper proposed a novel approach based on the partially-supervised alignment model, which regards identifying opinion relations as an alignment process. Then a graph-based co-ranking algorithm is exploited to estimate the confidence of each candidate. Finally candidates with higher confidence are extracted as opinion words. D. Ostrowski [5] proposed a methodology for the identification of topics associated with customers' sentiment using a Fisher Classification based approach towards sentiment analysis. By considering specific mutual information and word frequency distribution, topics are then identified within sentiment categories. The goal is to provide overall trends in sentiment along with associated subject matter. They demonstrated this methodology against data collected among a particular product line obtained from Twitter advanced search.

Saeideh Shahheidari et al. [6] described how to automatically collect Twitter corpus and built a simple sentiment classifier by utilizing the Naive Bayes model to determine the positive and negative sentiment of a tweet. Lastly they tested the classifier against a collection of users' opinions from five interesting domains of Twitter, i.e. news, finance, job, movies, and sport.

Marius Muja et al. [7] proposed a new algorithm for approximate nearest neighbor matching. For matching high dimensional features, they used the randomized k-d forest and a new algorithm proposed in this paper, the priority search k-means tree. One more new algorithm is proposed for matching binary features by searching multiple hierarchical clustering trees. This paper showed that the optimal nearest neighbor algorithm and its parameters depend on the data set characteristics and described an automated configuration procedure for finding the best algorithm to search a particular data set. In order to scale to very large data sets that would otherwise not fit in the memory of a single machine, they proposed a distributed nearest neighbor matching framework that can be used with any of the algorithms described in the paper.

III. PROPOSED SYSTEM

The activities involved in the process of opinion mining is illustrated in following diagram:



A. Data Gathering / Crawling

As E-commerce has propagated to a vast extend, more people are buying and selling more products online. The customer reviews that describe experiences with product and service use are becoming more important [1]. Potential customers want to know the opinions of existing customers to garner information about the products they plan to buy, and businesses want to find and analyze public or customer opinions of their products to establish future directions for improvement [2]. Customer reviews generally contain the product opinions of many customers expressed in various forms including natural language sentences.

Note that the target data set must be large enough to contain different patterns while remaining concise enough to be mined within an acceptable time limit. For feature - opinion extraction the customers' review is best suitable data. This can be crawled from any online shopping website.

B. Preprocessing / Data Scrubbing

User generated content (UGC) are mostly in unstructured format. UGC is also described as incomplete, noisy and inconsistent [8,9]. It is very difficult to process such unstructured data. It is necessary to apply certain preprocessing steps to the data so that it can be used for further analysis purpose. There are various preprocessing tasks such as, Data cleaning (identify and load in missing values, rectify noisy data, find and mark any outliers if present, and eliminate inconsistencies etc.), Data integration (combining multiple databases, multi-dimensional data, or files etc.), Data transformation (normalization and aggregation etc.), Data reduction (decrease the data size or data volume without affecting the analytical results etc.), Data discretization (it can be considered same as data reduction but replacing numerical values either with nominal ones or bucker of range). For opinion mining and analysis choosing appropriate tool for data preprocessing is essential. It creates a huge impact on the performance of entire system.

In the initial step of preprocessing for opinion mining one should eliminates the unnecessary content, such as tags, dates, and reviewer names, from the collected review data. Then, to extract noun phrases from the review data as feature candidates, NLProcessor [13, 14] is used to perform morphological analysis, including POS tagging.

C. Part of Speech Tagging

Part of Speech (POS) tagging is performing the task of dividing the text into words. They are treated as tokens. Then it searches for best suitable tag for each token. It tells us whether it is a noun, adjective or verb etc. Stanford POS Tagger can be used for this purpose. This speech tagger is using the model of maximum entropy. It enhances the performance by increasing the information sources used for tagging, incorporating more extensive treatment of capitalization for unfamiliar words. It also addresses the disambiguation of the tense forms of verbs, and focuses more on features for disambiguating part of words from prepositions and adverbs [15].

D. Dependence Analysis

Once the task of POS tagging is over, the next step is to extract the feature opinion pairs from the tagged tokens. This step is affected a lot by the nature of statement formation done by opinion writer. Few sentences are having clear subject and object part defined whereas few sentences are not having such clear separations [12]. So, it is not possible to extract opinions from all sentences. As a machine learning one can extract direct opinions only. The part of indirect opinion extraction is not considered here.

To extract candidate feature – opinion pairs different combination of dependences are used. The dependency parser finds out the different dependence relations between word pairs. For word w1 and word w2 , the dependence relationship is represented as relation_type(w1 , w2), in which w1 is called lead word and w2 is called dependent or modifier. The relationship relation_type(w1 , w2) can be either direct or indirect

E. Feature Opinion Extraction

As mentioned above for opinion extraction nsubj is the most important relationship. The process is feature – opinion pair extraction uses it. The relationship nsubj (nominal subject) is a noun phrase which is the syntactic subject of a clause. The governor of this relation might not always be a verb: when the verb is a popular verb, the root of the clause is the complement of the copular verb, which can be an adjective or noun[10,11, and 16]. These three cases namely governor as noun, governor as adjective and governor as verb should be treated separately. As per their governor one should change the extraction rules.

The extracted information component is represented by a triplet < f , m ,o >, where f stands for a feature generally articulated as a noun phrase, o stands for the opinion word which is generally articulated as adjective, and m is an adverb that acts as a modifier to represent the degree of expressiveness of the opinion.

For the present work only the noun – adjective pairs are considered. First extract the nsubj relationship. In line with above discussion then extract the respective feature from the relationship. Then find it’s corresponding opinion based on above mentioned rules. Once this step is over now the feature – opinion base is ready. The further step will find the performance of feature – opinion extraction process.

IV. PERFORMANCE MEASUREMENT

Evaluation of the experimental results can be performed using standard Information Retrieval (IR) metrics of Precision and Recall, these are defined in the below equations-

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

In these equations, TP indicates true positive, which is defined as the number of feature-opinion pairs that the system identifies correctly, FP indicates false positive which is defined as the number of feature-opinion pairs that are identified falsely by the system, and FN indicates false negatives which is the number of feature-opinion pairs that the system fails to identify. Based on the above criteria one can check for the performance of the system.

V. CONCLUSION

The proposed system works for feature – opinion extraction purpose. For small data set of customers review it shows proper results. Considering the UGC feature it gives less accuracy. To increase the performance of the system one should pay more attention on the nature of user’s writing style as well as the different dependence relationships generated by dependence parser.

REFERENCES

[1] W. Xindong, Z. Xingquan, W. Gong-Qing, and W. Ding, “Data Mining with Big Data”, IEEE Trans. Knowledge and Data Engineering, pp. 215-227, July 2013.

[2] B. Liu, “Sentiment Analysis and Opinion Mining”, Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167, May 2012.

[3] A. Das, “Opinion Extraction and Summarization from Text Documents in Bengali”, Ph. D. dissertation, Jadavpur University, Department of Computer Science & Engineering, Kolkata, December 2011

[4] K. Liu, X. Liheng, and J. Zhao, “Co-extracting Opinion Targets and OpinionWords from Online Reviews Based on the Word Alignment Model”, IEEE Trans. Knowledge and Data Engineering, vol. 6, no. 1, January 2013.

[5] D. Ostrowski, “Sentiment Mining within Social Media for Topic Identification”, in Proc. IEEE Fourth International Conference on Semantic Computing, pp. 394 – 401, 2010.

[6] S. Shahheidari, H. Dong, and M. N. R. Daud, “Twitter sentiment mining: A multi domain analysis”, in Proc. IEEE Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, pp. 144 – 149, 2013.

[7] M. Muja and D. Lowe, “Scalable Nearest Neighbor Algorithms for High Dimensional Data”, IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 2227 – 2240, vol. 36, no. 11, Nov.2014.

[8] A. Tatu, L. Zhang, E. Bertini, T. Schreck, D. Keim, S. Bremm, and T. Landesberger, “ClustNails: Visual Analysis of Subspace Clusters”, Tsinghua Science and Technology, ISSN l1007-0214 1l05/11, vol. 17, no. 4, Aug. 2012, pp. 419 – 428.

[9] N. Prendinger and M. Ishizuka, “SentiFul: Generating a Reliable Lexicon for Sentiment Analysis”, IEEE Trans. Affective Computing, vol. 2, no.1, pp. 22 – 36, June 2011.

[10] P. Balamurali, D. Manna, and P. Bhattacharyya, “Cross-Domain Sentiment Tagging Using Meta Classifier and a High Accuracy In-Domain Classifier”, in proc. of Eigth International Conference on Natural Language Processing ICON 2010.

[11] V. Subrahmanian and D. Reforgiato, “AVA: Adjective-Verb-Adverb Combinations for Sentiment Analysis”, IEEE Trans. Intelligent Systems, vol. 23, no. 4, pp. 43 – 50, July 2008.

[12] F. Peleja, J. Santos, and J. Magalhaes, “Ranking Linked-Entities in a Sentiment Graph”, in Proc. Int. joint Conf. Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM, pp. 118-125, 2014.

[13] S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, and X. He, “Interpreting the Public Sentiment Variations on Twitter”, IEEE Trans. on Knowledge and Data Engineering, vol. 6, no. 1, September 2012.

[14] B. Wang, Z. Xiao, Y. Liu, and Y. Xu, “SentiView: Sentiment Analysis and Visualization for Internet Popular Topics”, IEEE Trans. Human Machine Systems, pp. 620 – 630, Oct. 2013.

[15] M. Nagai, M. Ono, and R. Shibasaki, “Interoperability for Global Observation Data by Ontology Information”, Tinghua Science and Technology, ISSN 1007-0214 54/67 pp. 336-342, 2008.

[16] Z. Hai, K. Chang, J. Kim, and C. Yang, “Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance”, IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 3, pp. 623 – 629, March 2014.