Research Article

Opinion Extraction of Raw Data based on Classification Model

Pooja J,[1]* Sushma Ravindra[1]

Abstract: Sentiment analysis and opinion extraction is an important area to understand customer perception of the products. There are various existing approaches available to aid this purpose. Extracting the opinion from raw data is a challenging task, owing to the distributed data sources, inconsistent data formats. In this paper, based on the similarity of the text, opinion words and targets are extracted. Then the classification method adopted k-means algorithm.

Asian Journal of Engineering and Technology Innovation
Volume 4, Issue 7
Published on: 7/05/2016

Cite this article as: Pooja J, Sushma Ravindra. Opinion Extraction of Raw Data based on Classification Model. Asian Journal of Engineering and Technology Innovation , Vol 4(7): 110-115, 2016.

INTRODUCTION

Customer satisfaction is one of the most valuable assets for developing any enterprise customer relationship management. An organization always develops its strategic plan for their new products as well as for the cause of financial upliftment of the firm just with an aid of customer's opinion. Essentially, Customer Opinion can be defined as impartial perceptions borne by the customer about the level of personal experience with the product or the services offered to them by the service provider. Normally such forms of opinions are in the form of paragraph written by customer or it may also be certain form of recorded audio highlighting the speech of the customer [1]. Unfortunately, such conventional process of gathering customer's feedback is quite not only time consuming but also quite expensive in order to process it further for analysis. However, with the modernization of the communication and networking standards, now it has become feasible for gathering the customer's opinion in the form of text, which is not only fast in processing but also quite easy to understand. Although, analyzing the inference of the smaller version of the opinion shared by a single customer is not so difficult but problem escalates when massive volume of opinion in the form of text are accumulated. At present, there are various forms of text extraction techniques , that are already in use for the purpose of inferring the inner meaning of text-based opinion shared by the customer [2]. However, it was yet not found to be possible for exploring the best opinion inferring framework for assessing the behavior of the customer's opinion.

[1]Reva Institute of Technology and Management, Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Near Border Security Bustop, Bengaluru, Karnataka-560064, India.
E-mail: Pooja.jairam@gmail.com
*Corresponding author

In order to evaluate the text-based opinion, the present system utilizes the concept of data mining over the opinion shared in the form of text. The inputs for the opinion are possible to be gathered from different types of blogs, reviews in blogs, and finally a text document. As known that there are two categories of a simple sentence i.e. i) subjective and ii) objective. Normally, subjective sentence posses the significant amount of opinion to be extracted. Figure 1 shows the technique used for opinion mining. Initially, the system performs extractions of the data as well as domain knowledge and then it performs sentence processing. The next step is to perform analysis of the sentence in order to extract significant level of opinion for completing up the last stage i.e. data aggregation. Usage of opinion mining is widely found practicing in market review, product review etc from the customers[3]. The prime job of the opinion mining is to perform classification of the polarity of the investigated text gathered from opinion of a person in the form of positive, neutral, and negative.

LITERATURE SURVEY

In this section discusses about prior works carried out by different authors in the opinion mining domain.

Buche *et al* [4]. Have surveyed and analyzed various techniques that have been developed for the key tasks of opinion mining. The present research as concentrated in the domain of OM also known as SA due to presence of sheer volume of sentiments from web resources like discussion forums, survey sites as well as review blogs and sites available in digital form. One significant issues of SA of products reviewing is to generate a summary of sentiments on the basis of product features.

G.Vinodinin *et al* [5]. Presents a survey covering the techniques and methods in sentiment analysis and challenges

appear in the field. With the availability of rich opinion contents from the web such as discussion boards, reviewing forums blogs and news corpora people are willing to develop a system which can identify as well as classify sentiments that are represented in electronic text. A perfect mechanism for predicting sentiments will enable us to retrieve sentiments available in the internet and predict customer's preferences online, which can result in valued prize for economic or marketing research. Till date there is various challenges predominating research community such as feature based classification, sentiment classification and negation handling.

Singh *et al* [6] have attempted to review and evaluate the various techniques used for opinion and sentiment analysis. Decision making in organizational or even personal level are more often based on the search of other's opinion. With large resources of opinion such as reviews, discussion forum, micro-blogs, social media provide a huge collection of opinions. Such user generated opinion can be used for market beneficiation if the semantic orientations are deliberated.OM and SA is formalized and interpreting opinions and sentiments. Digital ecosystem has provided a huge space for large volume of opinions to be recorded.

Pak and Paroubek [7] have focused on using Twitter, the well known micro-blogging site for SA.Authors have illustrated how to automatically gather corpus for SA and OM purpose. Authors have also performed linguistic analysis of gathered corpus and demonstrated the discovered technique. By using the corpus authors constructed a sentiment classifier which is capable of determining positive, neutral as well as negative sentiments for document. Authors showed that through experimental evaluation their proposed technique is efficient and performed better than the earlier techniques. In this work authors have worked in English but the proposed techniques can also be used with any other language.

Severyn [et al] [8]. Defined a systematic technique to OM on comments in the YouTube by: i) Preparing classifier to predict opinion polarity and the comment type, ii) Proposed robust shallow syntactic design to enhance adaptability of the model. Authors used the tree kernel technology to perform automatic extraction and feature learning with good generalization power. By a extensive manual empirical evaluation of the YouTube comments corpus by the authors showed the high accuracy in classification and stressed the benefits of structural model in cross domain.

Al-kabi *et al* [9] have focused on evaluating Arabic social content. At present, Middle East is an field of prime political and social reform. Contents from the social media provide rich information to evaluate. Authors have constructed an OM and analysis tool to collect different kind of Arabic language (i.e. Standard or MSA, and colloquial). The tool generates output from the comments based on the priority. In addition tool will also determine comment or review whether it is subjective or objective, positive or negative and even strong or weak. Performance evaluation of the tool is done by applying it on domain-based Arabic reviews relative to general Arabic.

Qiu *et al* [10] Studied two significant issues, opinion lexicon expansion and opinion target extraction. Opinion target are entity and their feature on which opinions are conveyed. Authors have found several syntactic relations available to perform this task that link opinion words and target. The relations are discovered using dependency parser and then employ to expand the primary lexicon and for target extraction. Proposed mechanism is based on bootstrapping. Authors have termed it as double propagation since it propagates information between opinion word and opinion target. The prime advantage of the proposed method is that the it only requires a primary opinion lexicon to initiate bootstrapping process. This method is semi-supervised since it uses opinion words seeds. Authors performed the evaluation by comparing the proposed method with several other methods using standard product review test collection. It is found the proposed system outperforms the existing methods.

Woldarczak *et al* [11]. OM using social media is facing many challenges due to the presence of large volume of heterogeneity of data. Spam, fake opinions have emerged as a serious challenge. It is also effected by other challenges like language related challenges such as usage of jargon, slang ,special characters such as smiley's which are widely used in social media. Such challenges provide platform for research problem like calculating the power of social media in people's actions accepting opinion determining the online reputation of company. Nowadays opinion mining using social media as becomes an active area of research. Authors have discussed present state research and technologies being used in recent works.

Virmani *et al* [12]. Proposed SA in combination with opinion extraction, summarizing and tracking the student records. This work modified the existing algorithm in order to achieve the combined opinion about the student. The resulting opinion is represented as very high, moderate, high, very low and low. This work is based on the case studies wherein the teachers present their remarks about students and by applying the proposed SA algorithm the opinion is retrieved and represented.

Taylor *et al* [13] extended Bing Liu's Aspect-based OM technique and applied it to tourism domain. By this extension authors have also offered a strategy to consider a new alternative to identify consumer preference about tourism products in particular hotels and restaurants using sentiments available in internet. Using the reviews from Trip advisor a experiment is conducted to evaluate the proposed system. Proposed system is found to be very effective in analyzing sentiment orientation of pinions obtaining a precision and recall of 90%.

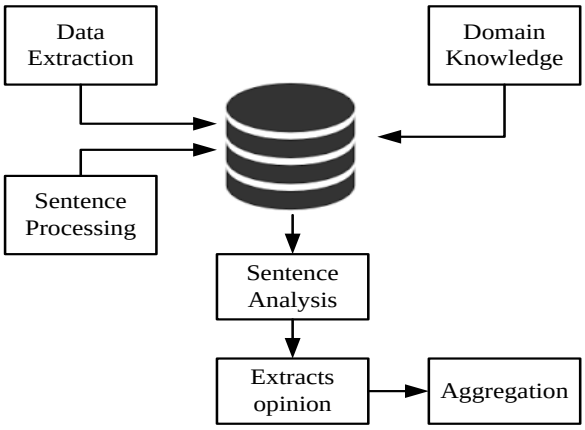

Figure 1: Opinion Mining Techniques

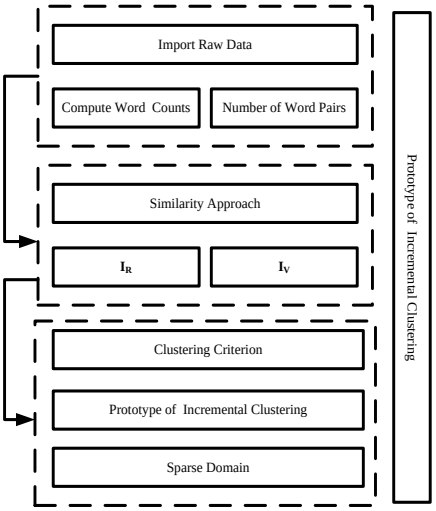

Figure 2: Architecture of Proposed system

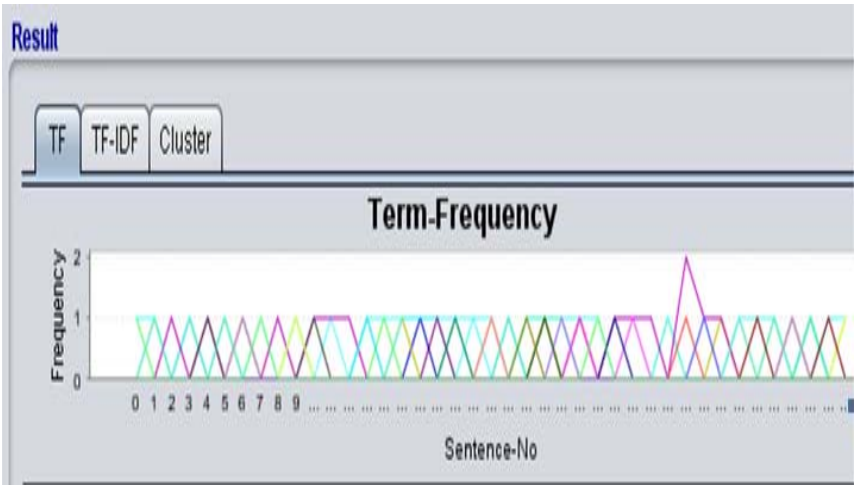

Figure 3. Graph representing the term frequency

Verma and Kiranjyothi [14]. Presented a review of all OM techniques which are used in extracting the opinions from social networking sites to identity the sentiments of online users. Authors stressed the importance of opinion conveyed in social network media in various domains. Different phases of OM like opinion extraction, analysis of opinion and

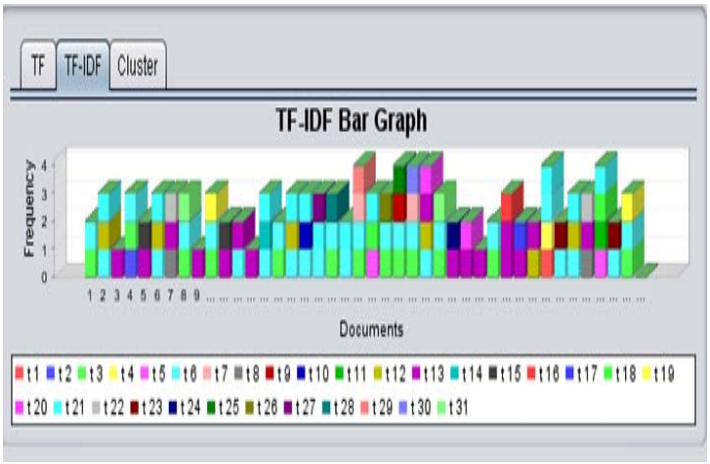

Figure 4: TF-IDF graphical representation

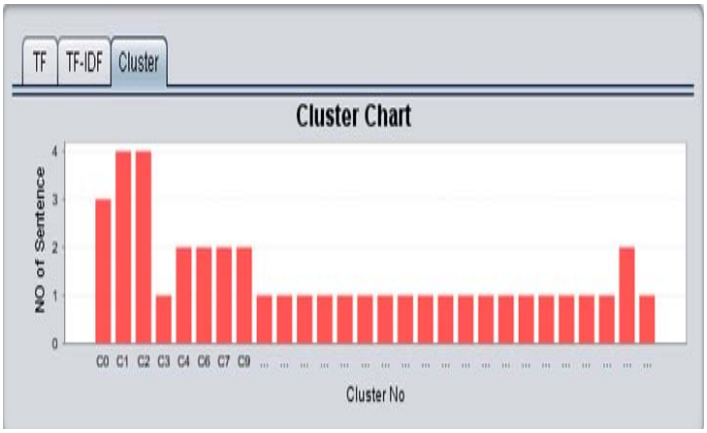

Figure 5: Clustering Graphical representation

classification are presented with their methods. Authors also suggest that an optimized framework for evaluation of the text in predicting the opinion on the basis of exact appraisal of user need to be set up.

Kessler *et al* [15]. Presented a task of predicting a ranking of products and introduced three potential sources for gold rankings: A sales ranking and expert based ranking have been used in the experiments in this paper. In addition, we discussed how to set up a crowd sourcing-based annotation of rankings. Authors have demonstrated early results how to use different opinion mining methods (dictionary-based, machine learning, comparison-based) to predict such rankings. In addition, we have presented experiments on how aspect-specific rankings can be used to measure the impact of that specific information on the ranking.

G. R. Krishna *et al* [16]. Has discussed, compared and analyzed 6 different OM algorithms. By means of PMI performance analysis is done. By experimental result it is seen that PAMM (Probabilistic aspect mining model) provides the best performance in compared to rest of 5 OM algorithms. By comparing different supervised topic modeling algorithms, PAMM provides distinctive characteristic which focuses on explanation aspects for single category.

Dhokrat *et al* [17]. Performed a brief review to cover the major challenges ,applications, stages and advantages of OM. Authors have reviewed few techniques like Naïve Bayes, SVM and Maximum Entropy which are often used in OM and SA.

PROBLEM DISCRIPTION

This section describes the problem associated with opinion mining. Through the extensive review of various work it is found that the opinion mining strategies are largely dependent on the lexical database. Since the present lexical database does not consider the semantics the analysis of such lexical database cannot be considered as accurate and such analysis can also produce more outliners. and one biggest disadvantage of the lexical database is that not all the words will be present in it. Certain time the words that are not present in the database produces inaccurate or even wrong analysis. Few other problems highlighted are the error generating from different aspects such as the typo error, wrong grammars, and even the difficulty of understanding the English language in various

parts of the world wherein a person may fail to understand other person's view due to the problem in one's language and its representation in expressing the issues or subject.

RESEARCH METHODOLOGY

This section illustrates the research methodology used to achieve the desired outcome of the project. The below Figure 2 shows the architecture of the proposed Classification Model which is further utilized to extract the knowledge based opinions from a raw data set.

The opinion mining process is performed on raw data containing text data. The raw data document is loaded by the data loader. After the data is loaded the data reader module performs the data reading by scrutiny line by line data of the document. The initial preprocessing operation starts with the segmentation primarily with line segmentation by dividing the whole documents to readable sentence. Segmentation is also responsible for removal of stop words. After the sentence segmentation word segmentation is achieved. On completion of word segmentation a wordlist is prepared which is vector defining word in the sentence. On completion of preprocessing, different operations such as similarity approach, computation of computation IF and IDF is performed, Here the IF and IDF terms are represented as IR and Iv which is followed by calculating the centroid and cosine similarity. Finally the classification is based on the opinion target. This analysis also makes use of available lexical data base such as WordNet which provides different features like analysis a word into noun, verb or adjective in order to increase the accuracy of the opinion mining. In the following section a detail explanation of the proposed opinion mining process is performed. The algorithm used for this process is illustrated.

IMPLEMENTATION AND ALGORITHIM

This section illustrates the algorithm used for the implementation in achieving the desired outcome.

Table 1: Algorithm Description

Input : Raw data file (.Txt)
Output : Opinion words and opinion Targets
Start:
1) Load the Data file
2)Initialize Data Reader
3) Perform Segmentation
4)Sentence Segmentation
{
Buffer reader →Line to line data read
Initiate trim operation }
5) Initiate word segmentation
{ Wsegmenation→Wordlist}
6) Similarity approach {
PreprocessedData→ProcessedKnowledgeClass
(Adjectives,Verbs,Noun)
If (Lexdata ϵ Wint)
{
Wint available
{
If (Wint ϵ Noun)
{
Wint = OT
}
Else if
{
Wint ϵ Adjective
Remove Wint from adjective.
Add Wint into OW
}
Else
{
Wint =RData
}
7) Check IF and IDF
{
TF= occurrence of Wint/ total number of occurrence Wtotal
IDF= log (total Tsize / occurrence of the word within the complete document).
}
8) Calculation of centroid
{
Centroid= Dot product/ Euclidean dist.
}
9) Calculation of Cosine similarity
$\cos\theta = n\left(\frac{A}{\|A\|}\frac{B}{\|B\|}\right)$
$\cos\theta = n\left\{\frac{A}{\|A\|}\frac{B}{\|B\|}\right\}$
Where n is the Multiple points, and A and B are Words
}
10) Classification process
{
Tsen ϵ C
Where C= (C1, C2,……..Cn)
} End

RESULT AND DISCUSSION

This section discusses the result obtained and the evaluation of the proposed system. Here the evaluation is graphically expressed for different operations.

This (Figure 3) graph illustrates the term frequency in a text. It gives the presence of a specific word in a sentence. In the graph different words are represented with different colors. Another evaluation performed is the calculation of the IF and IDF, which are crucial in the opinion mining process.

TF and IDF are significant since they help to evaluate the occurrence of the words and can be used as the weighing factor for analysis.

Figure 5 shows the clustering graph. Clustering is a key process in opinion mining as it is used for the classification by which the analysis of the query is performed.

CONCLUSION

Opinion extraction from the raw data using the cosine similarity and k-means approach provides promising results. In order to improve the accuracy on classification based on clustering, we have used n point cosine similarity in compared to the two point cosine similarity used in conventional classification.

REFERENCES AND NOTES

1. Bo Pang, Lillian Lee,"Opinion Mining and Sentimental Analysis",Now Publishers Inc, 2008 - Computers - 137 pages
2. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. 2010.
3. Jin, Wei, Hung Hay Ho, and Rohini K. Srihari. "OpinionMiner: a novel machine learning system for web opinion mining and extraction."Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.
4. A. Buche, M. B. Chandak, A. Zadgaonkar, "Opinion mining and Analysis : A survey", International Journal on Natural Language Computing (IJNLC) vol.2,No.3,June 2013G.
5. Vinodini and RM Chandrasekaran,"Sentiment Analysis and Opinion Mining: A survey",International journal of Advanced Research in computer science and software engineering,vol 2,Issue 6,june 2012.
6. P. K. Singh, M. S. Hussain, "Methodological study of opinion mining and Sentiment analysis techniques", International Journal on soft computing (IJSC) vol 5,no 1. feb 2014.
7. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. 2010.
8. Severyn, Aliaksei, et al. "Opinion Mining on YouTube." ACL (1). 2014.
9. Al-Kabi, Mohammed N., et al. "Opinion mining and analysis for arabic language." IJACSA) International Journal of Advanced Computer Science and Applications 5.5 (2014): 181-195.
10. Qiu, Guang, et al. "Opinion word expansion and target extraction through double propagation." Computational linguistics 37.1 (2011): 9-27.
11. Wlodarczak, Peter and Ally, Mustafa and Soar, Jeffrey, Opinion Mining in Social Big Data (February 15, 2015)
12. Virmani, Deepali, Vikrant Malhotra, and RidhiTyagi. "Sentiment Analysis Using Collaborated Opinion Mining." arXiv preprint arXiv: 1401.2618 (2014).
13. Marrese-Taylor, Edison, et al. "Identifying customer preferences about tourism products using an aspect-based opinion mining approach." Procedia Computer Science 22 (2013): 182-191.
14. R. Verma, Dr. Kiranjyoti,"Opinion Mining and Analysis of the Techniques for User Generated Content (UGC)", "International Journal of Advanced Research in Computer Science and Software Engineering",vol 5,may 2015.
15. W. Kessler, R. Klinger, and J. Kuhn, "Towards Opinion Mining from Reviews for the Prediction of Product Rankings",WASSA 2015,Portugal,Sept-2015
16. G. R. Krishna, S. Kavitha, S. Yamini, A. Rekha, "Analysis of Various Opinion Mining Algorithms",International Journal of Computer Trends and Technology (IJCTT),April-2015.
17. A. Dhokrat, S. Khillare, C. N. Mahender, "Review on Techniques and tools used for Opinion Mining",International Journal Of Computer Applications Technology and Research, Volume 4, 2015.