ROM - Review Opinion Mining a Novelized Framework

K. Vivekanandan[1*], V.L. Helen Josephine[2*]

[1]*Bharathiar University, Coimbatore, Tamilnadu, India*

[2]*School of CSE, Bharathiar University, Coimbatore, Tamilnadu, India*

Abstract—Today, as a result of the global internet viewers increased rapidly, consumers are more focused than ever on searching the best product and the best prices. Consequently, e-commerce corporations also invested their time, money and efforts to know the feedback and comments about their products. That would help the corporations to modernize their product at low prices, which in turn help them to extend and prosper in their business. Customer / Product review is an evaluation of the product performance and comment on the reliability and whether or not the product delivers on these promises. Now-a-days, online reviews are the recent media world-of-mouth, they are enormously influential and may have an enormous effect on however business is perceived. Since, overwhelming information on one product is available in the form of review, individuals or corporation finds very difficult to analyse each and every review to extract knowledge from that pool of unstructured data. So, to analyse and to extract knowledge from these large amounts of data automatic method must be developed. This paper describes the ROM framework for developing such an automatic method to mine the opinion from the online product reviews.

Keywords—Opinion Mining, Sentiment Analysis, Framework for Opinion Mining

I. INTRODUCTION

Opinion Mining places a vital role in business for both consumers and producers. Now a days, with the expansion of internet, online shopping is more common in day to day life. The consumer wants to select and buy a particular product from the voluminous collection of product. In online shopping it is not possible to get help from the marketing person about the product they wish to buy, but rather customer depends on product review and comments which was given by other consumer already bought and used the particular product. Manufacturers also very much interested to know about the consumers' feedback or comment to extend their business.

Due to the rapid growth of the internet over the last decade, opinions will currently be found nearly everywhere – online customer reviews, forum discussions, blogs, social networking sites like Facebook, Orkut and Twitter, news-portal, e-commerce sites, etc. Producers wish to research the feedback of consumers concerning their product to create future selections. But they faced to wade through many on-line reviews in order to make an informed product choice. So, analysing customers' opinion and their reviews and comments are very important.

Sentiment analysis, also called opinion mining, is the field of computational study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes[1]. By using the opinion mining, automatic method must be established to analyse and extract knowledge from the volume of online product review.

In general, opinion mining has been explored mainly at three levels [2].

a) Document level

b) Sentence level

c) Entity and feature / aspect level.

Both the document level and the sentence level analyses do not determine which aspect people liked and did not like. Aspect level achieves fragile analysis and yielding the best result for the problem of review mining. Aspect level was earlier called *feature level (feature based opinion mining and summarization).[3]* It is based on the inspiration that, an opinion contains with a *sentiment* (positive or negative) and a *target* (of opinion). An opinion without its target being identified is of limited use. Realizing the significance of opinion targets assists the researchers to understand the sentiment analysis problem with fine-granularity.

Example :

I bought Fujifilm X-E2 digital camera two weeks ago. The picture quality is first-class. The battery life is also too long. But, quick menu button a little cumbersome to reach.

Corresponding Author: V.L. Helen Josephine,
helenjose.cbe@gmail.com

- Entity : Fujifilm X-E2
- Features / Aspect
    - Picture Quality
    - Battery life
    - Quick menu button
- Opinion
    - First-class
    - Too long
    - Cumbersome
- Polarity identification
    - First-class – Positive
    - Too long – Negative
    - Cumbersome – Negative

This paper deals about the ROM framework for developing such an automatic method to extract knowledge from the online product reviews. The framework is developed in such a way that the unstructured data (online review about the particular product) are pre-processed, identifying the features of the product, identifying the opinions related to the product features, determining the polarity of the opinions and summarizing all the opinions based on their polarity.

The rest of this paper is ordered as follows: In section 2 motivation is discussed and then in section 3 proposed framework ROM and pseudo code is explained. Finally, this paper is concluded in section 4 with future work.

II. MOTIVATION

A customer who wants to buy a new product looks for reviews and comments for finding the best product and the best prices. Most of the time, customers who just brought the product writes about the experiences, product performances, whether or not the product delivers on the promise. At the same time, product manufacturers get the feedback from the customer to improve their product and adjust marketing strategies.

But in the real scenario, it is difficult to read, analyse and grasp the entire review comments. If the customer who wants to buy / manufactures missed to go through some of the reviews, then biased assessment about the product may be generated and eventually it leads to take the wrong decision. So, to analyse and extract knowledge from these large amounts of data, some form of summary of opinions is required to assist the consumer to take a correct decisions.
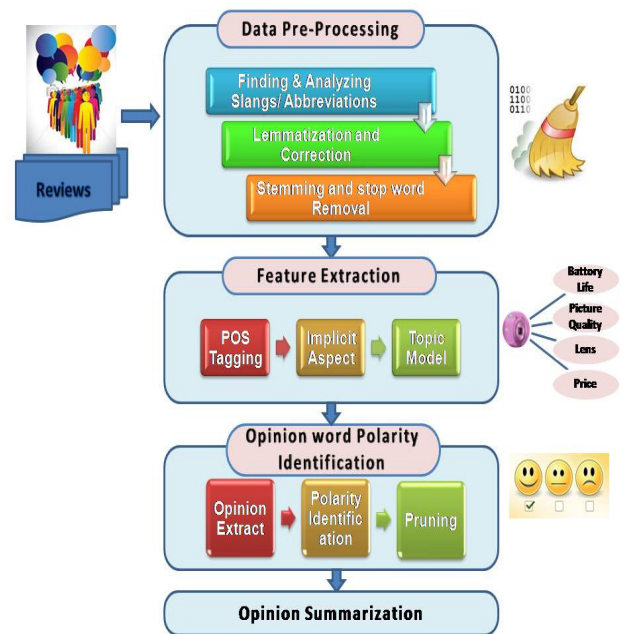
III. PROPOSED FRAMEWORK - ROM

Figure1 shows the proposed ROM framework to analyse and summarize the opinion from the unstructured data namely product reviews. The principle of the framework is as follows

(i)   To identify incomplete, noisy and inconsistent data and clean the dataset.
(ii)   To identify the features of the product
(iii)   To identify the opinions word and determine the polarity
(iv)   To determine the polarity of opinions
(v)   To summarize feature based opinions on their polarity

The steps are as follows,

- A. Data pre-processing
- B. Feature / Aspect extraction
- C. Opinion identification
- D. Opinion's polarity Identification
- E. Feature based summarization

A. Data pre-processing

The function of this step is to identify and eliminate the incomplete noise, inconsistent data in the unstructured data set. Data must be pre-processed to achieve quality data. Quality decisions must be based on the quality data. In this step, to achieve quality data the following pre-processing activities are carried out [4].

- a) Spelling corrections using a dictionary
- b) Find the slangs, special characters and replace the appropriate word
- c) Lemmatization and stop words removals and
- d) Expanding abbreviations.

Pseudo code for ROM Framework

1. *For each review do*
    function pre-process(reviews)
    {

        Specll check & correction
        Find slags and replace appropriate word
        Expand abbreviations
        Remove stopwords
        Lemmatization
        Find special character and replace opinion word

    }
2. *function feature_extraction(refined reviews)*
    {

        POS Tagging
        Parsing
        Extract implicit aspect by enhanced_featurextractor

    }
3. *function opinion_polarity_identification(refined reviews)*
    {

        Determine opinion polarity by
                enhanced_opinion_polarity_identfier.

                Genertate_feature_opinion_ pair
                  return(feature, opinion pair)
                            }

4. *function feature_based_summarization(feature, opinion polarity pair)*
    {

        Summarize feature based opinion by
                enhanced_feature_based_summary
        write the summarization result to the outputstream

    }

B. Feature / Aspect exttaction

Feature/Aspect extraction involves identifying salient features within the entity. [5] In the above example picture quality, battery life and quick menu button are considered as the salient features/ aspect of the Fujifilm X-E2 digital camera. To generate opinion about the digital camera, common aspects are 'picture quality', 'battery life', 'quick menu button'. The function of this step is to extract common features of the entity.

There are two types of aspects (i) explicit aspect (ii) implicit aspect. POS tagging and parsing are very effective for finding explicit aspects. In Example 1 'picture quality', 'battery life', 'quick menu button', are noun phrases known as explicit aspects easily extracted by POS tagging and parsing techniques.

But consider following example 2,
Example 2 : The refrigerator is not good to make ice.
          Entity : refrigerator
          Feature : ice making performance

In the above example, the feature is not explicitly specified in the text, the word 'mark' is a verb determines the feature 'ice making performance' of the refrigerator. This is because features are not always nouns, sometimes verb can be used to extract the feature [6].

C. Opinion identification

The function of this step opinion identification is to extract all the opinion words in each review. Different people may have used different opinion word to describe their views about the features of the product[7]. In the above example 1, First-class, too long and cumbersome are the different words to describe the opinion about the features. The outcome of this step opinion identification, would facilitate grasping the common sentiments on the features. Enhanced_feature_xtractor algorithm to be proposed to extract features from the reviews.

D. Opinion's polarity identification

This step is the communication of the previous step, opinion word's polarity (negative, positive) is identified. Then finding the feature-opinion pair in order to generate feature based summarisation. Many literatures are concentrated to detect opinion word's polarity. Opinion word classification in itself is an energetic research area [8]. Mainly three different approaches are used to detect opinion words' polarity.

   a) Supervised learning approach
   b) Lexicon-Based approach
   c) Integrated approach (Lexicon based along with Supervised learning based)

Based on the integrated approach enhance_opinion_polairty_identifier algorithm and generate_feature_opinion_pair algorithm will be proposed.

F. Feature based Summarization

By using the outcome of opinion polarity identification step, future based summarization is will be generated in an effective, comprehensive format in the output file. This effort is associated with the problem of multi-document summarization. Each review is considered as the multi document. According to [9] the following methods are used for multi-document summarization.

a) Centroid-Based Method [10]
b) Graph-based Methods [11] [12]
c) Latent Semantic Analysis (LSA) [13]
d) Non_Nagative matrix Factorization (NMF)[14]

e)   CRF-Based Method [15]
f)   Hidden Markov Model (HMM) Method[16],[17]

Most of the above methods are intended for traditional document summarization tasks. They above methods are not used feature opinion pair for the summarization.   The proposed algorithm enhanced_feature_based_summary is combining both positive approaches of traditional document summarization with the feature opinion pair.

IV.   CONCLUSION

This paper mainly focuses about the ROM framework for review opinion mining. This framework involves the steps which represent the different activities that are performed to extract knowledge from the unstructured data. Different algorithm to be proposed to identify in the steps A to E which helps to mine the opinion from the large amount of data.

By using the ROM framework, digesting the information available in large numbers of products review corpus has become fast and efficient. In future work, we will aim to develop appropriate algorithms for each step A to E with more accuracy.

AKNOWLEDGMENT

REFERENCE

[1]   Liu, B. 2010. Sentiment analysis and subjectivity. In Handbook of Natural Language Processing, Second Edition, N. Indurkhya and F. J. Damerau, Eds. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

[2]   Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retr. 2, 1-2, 1–135.

[3]   Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[4]   Hemalatha, G.P. Saradhi Varma, A. Govardan, Preprocessing the Informal Text for efficient Sentiment Analysis, IJETTCS, Volume 1, Issue 2, July-August 2012, ISSN: 2278-6856

[5]   Kushal   Bafna,   Durga   Toshniwal,   Feature   Based Summarization of Customers' Reviews of Online Products, Elsevier Procedia Computer Science 22(2013) 142-151.

[6]   Alexandra Blahur, Mijali Kabadjov, Josef Steinberger, Ralf Steinberger, Andres Montoyo, Challenges and solutions in the opinion summarization of user-generated content, Springer Science, J Intell Inf Syst (2012) 39: 375-398

[7]   Bakhtawar Seerat, Farouque Azam, Opinion Mining : Issues and Challengers (A Survey), IJCA, (0975-8887), Volume 48 – No. 9 July 2012.

[8]   Kim, Hyun Duk, Ganesan Kavita A., Sondhi Parikshit, and Zhai ChengXiang,   Comprehensive Review on Opinion Summarization, 2011.

[9]   Dingding Wang, Shenghuo Zhu, Tao Li, SumView: A Web-based engine for summarizing product reviews and customer opinions, Elsevier, Expert Systems with Applications 40 (2013) 27-33

[10]   Radev, D., Jing, H., Stys, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing and Management, 919–938.

[11]   Mihalcea, R., & Tarau, P. (2005). A language independent algorithm for single and multiple document summarization. In Proceedings of IJCNLP 2005.

[12]   Wan, X., Yang, J., & Xiao, J. (2007). Manifold-ranking based topic-focused multi-document summarization. In Proceedings of IJCAI (pp. 2903–2908).

[13]   Gong, Y., & Liu, X. (2001). Generic text summarization using relevance   measure   and   latent   semantic   analysis.   In Proceedings of SIGIR (pp. 75–95).

[14]   Li, T., & Ding, C. (2006). The relationships among various nonnegative matrix factorization methods for clustering. In Proceedings of IEEE international conference on data mining (pp. 362–371).

[15]   Shen, D., Sun, J.-T., Li, H., Yang, Q., & Chen, Z. (2007). Document summarization using conditional random fields. In Proceedings of IJCAI (pp. 2862–2867).

[16]   Conroy, J., & O'Leary, D. (2001). Text summarization via hidden markov models. In Proceedings of SIGIR (pp. 406–407).

[17]   Dhanashri Chafale and Amit Pimpalkar, "Review on Developing Corpora for Sentiment Analysis Using Plutchik's Wheel of Emotions with Fuzzy Logic", International Journal of Computer Sciences and Engineering, Volume-02, Issue-10, Page No (14-18), Oct -2014, E-ISSN: 2347-2693

AUTHORS PROFILE

Dr.K.Vivekanandan obtained Ph.D in Computer Science in 1996. He has produced 10 Ph.D candidates and 7 M.Phil candidates. He has published   articles   in   16   national   and international journals. His area of interest includes data mining and information systems.

Mrs. V.L. Helen Josephine received the Master of Computer Applications in 1996. She is pursuring Ph.D. Degree in the Department of Computer Science and Engineering,   Bharathiar   University, Coimbatore, India.   Her research interest includes Web Mining, Opinion mining and sentiment analysis. She is a life member of ISTE.