Volume 6, Issue 1, January 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Unified Classification Approach to the Online Review Selection using Micro- reviews

Jananee. M, Devi Selvam
Department of Computer Science and Engineering,
Sri Shakthi Institute of Engineering and Technology, Tamilnadu,
India

Abstract -Micro-reviews are a considered as a source of review content in the social media. In this work, we study the micro-reviews, and use them to solve the problem of review selection. We introduce a novel formulation of review selection, to ensure efficiency and increase the coverage, which leads to novel coverage problems. The mechanism for helping the users with comprehensive set of reviews which satisfies their need is critical. So, here we consider approximation and heuristic algorithms, and study them experimentally, in order to demonstrate quantitatively and qualitatively the benefits of our approach. We also propose an Integer Linear Programming (ILP) formulation and an optimal algorithm. This helps us to quantify the approximation quality of the greedy heuristics. Experimental results explain the performance of the system against the State of art approach in terms of precision and recall.

Keywords: Data classification, online social platform, opinion mining, micro-review

I. INTRODUCTION

A review is basically an assessment of a publication, service, or company or an event or performance. An author may review current events, trends, or items in the news. A collection of reviews is called a review.A review site is a website on which reviews can be posted about people, businesses, products, or services. The review sites which were used earlier include Epinions.com and Amazon.com.Web sources provide more amount of review content. Readers are overloaded by the information provided on the web and find it difficult to extract the required information. The reviews posted on web are worsened by length and may not be relevant to the user. A list of reviews may not represent the different viewpoints such as positive and negative opinion of the item being reviewed. The readers have to filter the review for extracting the set of desired opinion. Mining the high quality reviews is a difficult task.

Presently, due to the growth of social networks and micro blogging services, a new type of online review content is available. This is known as the "micro-reviews". There are many sources for micro-reviews. The micro-review sites are the sites which allow the reviewers to post the reviews with the maximum length of 200 characters. It is an alternative source of content for the readers to search the desired reviews. The information provided must be compact and comprehensive.

Some of the importance of micro-reviews is

1) It is concise and distilled because of the length and helps in identifying the desired information.
2) It can be written only after the user has checked it and hence it is considered to be authentic.

Micro-reviews and reviews are complement to each other. However, the micro-review deals with only the comprehensive information, it cannot cover all the aspects of an item with different viewpoints. Combining both the reviewing approaches, we can focus on the true requirements of the user.

Some of the applications of the reviewshave been described below:-

- *Advertising:*
  Each review which is posted by the customer on the internet is a free advertisement to the business. The online reviews can cover many businesses which also includes the small business.
- *Peer recommendations:*
  Most of the customers generally trust the peer recommendations more than they trust the advertisements. A powerful form of marketing is when the online reviews recommend the business or the product.
- *Constructive criticism and suggestion:*
  The online reviews are used to raise concern and suggestions for improving the business. It provides an opportunity to resolve the customer problems and to improve the practices.
- *Closer relationship with customers:*
  The online review sites provide a chance to develop a closer relationship with the customers. It offers an opportunity to reply for the positive and also the negative opinions of the customer.

The prior worksare based on the coverage problem where collection of reviews and tips for the product manufactured is considered. The problem is to select the small number of reviews that cover the tips.[1] The selected

reviews must also cover the several different opinions of the product. This application was efficiently used in several websites and mobile applications which are in need of displaying the smaller number of reviews. For example, the review sites which make use of this technique are Yelp, Twitter, Google Local and many more.

The task of selecting the comprehensive set of reviews has been studied in the past. This was meant for the online sites like Amazon.com, TripAdvisor.com, where several hundreds of reviews are being posted.[7] Few of these reviews may be fraudulent, redundant or uninformative. To solve this problem the reviews are ordered and a score for the ordered list of reviews is made. The drawback is that the ordered list of reviews does not represent all the different opinions (positive and negative) of the product. The top reviews in the order would generally represent the single view point or opinion of the product.[8]

Our work focuses on matching the reviews with the micro-reviews. First, the unwanted words are removed from the sentences and then we need to identify the type of opinion which is generated from the tip, it can be positive, negative or neutral. The subset of reviews which match with micro-reviews are chosen and the optimal solution is retrieved from this set of statements.

II. BACKGROUND

The growth of online sites and the review content is tremendous in the present time. The fact is that the reviews are highly diverse and often unnecessarily verbose. Selecting the appropriate reviews is difficult for the users, since there are huge numbers of reviews available on the online sites. Micro-reviews are emerging as a new type of online review content in the social media. Micro-reviews are posted by users of check-in services such as Foursquare. They are concise (up to 200 characters long) and highly focused, in contrast to the comprehensive and verbose reviews. In this paper, a novel mining problem is proposed, which combines the two disparate sources of review content. Specifically, we use coverage of micro-reviews as an objective for selecting a set of reviews that cover efficiently the salient aspects of an entity. This approach consists of a two-step process: matching review sentences to micro-reviews, and selecting a small set of reviews that cover as many micro- reviews as possible, with few sentences. The objective is formulated as a combinatorial optimization problem, and shows how to derive an optimal solution using Integer Linear Programming. It also proposes an efficient heuristic algorithm that approximates the optimal solution.[1]

III. LITERATURE REVIEW

A. Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinion

In this literature, an ultra-concise summary of opinions is generated. The summarization is helpful for users to filter the relevant reviews from the reviews which are posted on the web. The summaries must be representative to the key opinions and must be in a format so that users can understand it easily. The optimization problem is to search the set of precise and non-redundant text that represents the key opinions in the reviews. The task of generating the textual opinion summaries is difficult. A micropinion is generally a short phrase that summarizes a key opinion in text. The main objective of this paper is to optimize the representativeness and readability in-order to ensure that the summaries reflect the opinions of the original text and it is also well-formed. Heuristic algorithms are used to solve this problem which uses the steps like : Generating seed bigrams, scored n-grams and micropinion summary. These methods are used to reduce the redundancies in the document. This shortlists the set of words used to generate the n-grams. It is based on the idea that the words that are not frequent in the selected review document is not considered as a good candidate which can be included in the micropinion summary. The high frequency words are considered as unigrams. Each unigram is combined with the other unigram to form bigrams. The depth first search is used for generating the candidate words. This approach is general and lightweight and does not require any domain knowledge. It can be used in other domains and also in other languages.[2]

B. Exploring temporal effects for location recommendation on location-based social network

This literature is based on the Location Based Social Networks(LBSNs).This work provides a point of interest(POI) to the user according to the required preferences and helps the users to explore new areas of the city. The LBSN provides large scale check –in data which is used to describe the user's mobile behavior. The behavior is described according to the spatial, temporal and social aspects. It analyses the temporal properties and proposes a location recommendation framework. This work is based on two types of temporal properties which are described below:

a) Non-uniformness:
   A user has different check-in preferences at different hours of the day.
b) Consecutiveness:
   A user has many similar check-in locations in consecutive hours than in non-consecutive hours.

The main objective of this literature is to leverage the temporal properties for location recommendation. This is accomplished using the low-rank matrix factorization. It also introduces few temporal aggregation strategies. The results show the power of temporal effects in capturing a user's behavior and improvement in the location recommendation framework compared to the other works.[3]

C. Twitter sentiment analysis: The good the bad and the omg

This paper discusses about the Twitter sentiment analysis. It is considered as very different task. It makes use of the linguistic features for detecting the sentiments of the messages posted on the internet. It evaluates how the lexical

resources which already exist are useful in the sentiment analysis. The results shows that the part of speech feature is not useful for the sentiment analysis, instead hashtags can be used to collect the training data. The data collected must have both positive and negative opinions. There are three types of corpora used in this paper which are hashed data set, emoticon data set and annotated data set.

a) The hashed data set is used for the development and training process. The top positive, negative and neutral hashtags are used to create the hash dataset.
b) The emoticon data set is used for identifying the emotions of the messages. The messages which express both positive and negative opinions are neglected.
c) The annotated data set is used for the evaluation process. The data selected for evaluation is based on certain subjects and the label of each tweet would reflect its sentiment.

The data processing is used to accomplish the goal of this paper. It uses three steps, which includes: tokenization, normalization and part-of-speech. A binary file is created in order to capture the emoticons like positive, negative and neutral. [4]

D. Designing novel review ranking systems: Predicting the usefulness and impact of reviews

This literature discusses about the ability of the user to publish the information related to a product in order to create an active electronic community. The user who wants to buy a product would read the reviews and decide whether to buy a product or not. Due to the huge amount of reviews posted for every single product, the users find it difficult to choose the best review and also the quality of the product based on the posted reviews. The product manufacturers must also read the reviews for identifying the customer's need and also must examine the contents of each review posted by the customer. To address this problem two ranking mechanisms are used in this literature:

a) Customer-oriented ranking mechanism:
   It is used to rank the reviews according to the expected helpfulness.
b) Manufacturer-oriented ranking mechanism:
   It is used to rank the reviews according to the expected effects on sales.

This paper proposes a fact that reviews which include a combination of subjective and objective elements are more informative.[5]

E. Mining and summarizing customer reviews

This literature is based on the fact that the reviews must be mined and summarized according to the needs of the customer. The manufacturers who sell products through the web have the need to know the review about the product from the customer. The number of customer reviews for products have increased due to the growth of e-commerce. For any product which is popular among the customers have more number of reviews. So, it is a difficult task for a customer to read all the reviews and to make a proper decision to buy or reject a product. It is also hard for the manufacturers to maintain the product according to the customer's opinions. The task of summarization used in this work is different from the traditional work which involves the following steps:

1) Features of the product are identified on which the customers have expressed their opinion.
2) For each feature the review sentence is identified.
3) Using the discovered information a summary is produced.

It mines the features of the product for which the customers have given their review and decides whether it is positive or not. [6]

Table I. Comparison Table for Review on Literature

Paper Title	Description	Advantages	Disadvantages
1. Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinion	An ultra-concise summary of opinion is generated. The summarization is helpful to filter the relevant reviews.	Data Summarization technique yields the output in short span of time.	The content is considered as a list of short phrases. The performance of summarization is not sensitive to the settings of the parameter.
2. Exploring temporal effects for location recommendation on location-based social network	This work provides a point of interest(POI) to the user according to the required preferences and helps the users to explore new areas of the city.	Location based recommendation has the influence of both social influence and spatial information.	Most of the check-in services are not considered as the source of reviews but are considered as a location based social network(LBSN). The temporal patterns are used to investigate the other patterns.

3. Twitter sentiment analysis: The good the bad and the omg	It was proposed to investigate the utility of the linguistic features for detecting the sentiment of Twitter messages.	It is used to extract the linguistic features in the micro review as wells in the large reviews.	The micro-blogging services are used to analyze the opinion. Most of the opinions are based on more general concepts rather than specific entities. Difficult to separate reviews from other content.
4. Designing novel review ranking systems: Predicting the usefulness and impact of reviews	This literature discusses about the ability of the user to publish the information related to a product in order to create an active electronic community.	The sentiment of the text in the review is analyzed to show how the review affects the product sales and the extent to which the reviews are informative.	Finding the good set of reviews and checking the quality is very difficult. It makes use of the classification approach or supervised regression for selecting the top reviews, which is not much efficient.
5. Mining and summarizing customer reviews	This literature is based on the fact that the reviews must be mined and summarized according to the needs of the customer. A task of summarization is used to identify the relevant reviews.	It mines the features of the product for which the customers have given their review and decides whether it is positive or not.	A complication occurs in feature tagging as there are explicit or implicit features in the review sentence. Making decisions on opinions in reviews are subjective.

IV. PROPOSED SYSTEM

Due to the growth of online business websites, the idea of reviews was implemented. More number of people have started purchasing the products online and the number of reviews for each product is getting increased. These reviews are lengthy and the verbosity is also high. Processing the large set of reviews is time consuming. In order to overcome this problem, we propose the online review selection using the micro-reviews.

Micro-reviews are an alternative source of the review. The maximum review length is 200 characters and hence the processing time would be lesser. The main objective of our work is to match the reviews with the micro-reviews and derive a subset of the reviews. The optimization problem is to find out the set of reviews which are relevant to the user. This is accomplished using the heuristic algorithms. The result which has simulated shows the better efficiencyfor the classifiers.

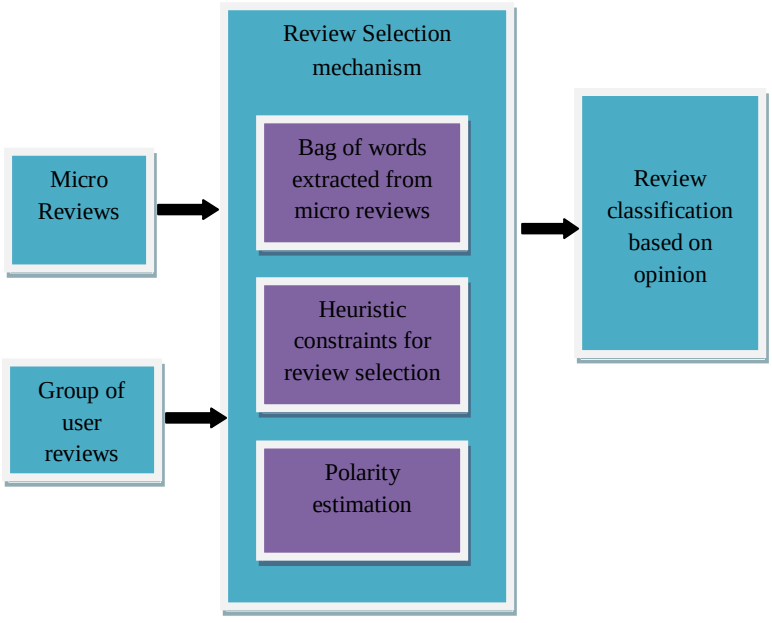

Figure 1: Architecture Diagram

V. METHODS

A. Bag of words construction

The important sentence and the tip are considered as bags of words. It is extracted from the reviews of the products using parts of speech tagging or stop word removal and sentence splitting mechanism.The data extracted shares a substantial subset of textual content then the data is assumedto convey a similar meaning.

B. Concept and opinion generation

Sentence and a tip may discuss the same concept, but use different words according to the situation.In this process, we mustdetermine an approximation bound for the two variants of the efficiency function in the micro review and review. Against the review selection, there exist important aspects to determine positive and negative opinionated sentences which are often used to extract comparable sentences from each set of opinions and generate a comparative summary containing a set of contrastive sentence pairs.

C. Selecting of Subset of Reviews

Some reviews may have high coverage, but at the same time they are too verbose. Containing many sentences are not relevant to any tip. We would like to avoid such reviews in our selection, so we introduce the concept of efficiency. It means that if a sentence s and a tip t are matched, then we say that s covers t. We will say that a review R covers a tip t if there is a sentence s in R that is matched to the tip t. Given the collection of reviews R and the collection of tips T , and the matching function F, we define for each review R the set of tips T R that are covered by at least one sentence of review R.

D. Generating set of the reviews as heuristics

It is well known that due to the sub modularity property of the coverage function, the greedy algorithm that always selects the review whose addition maximizes the coverage produces a solution with approximation ratio. The intuition is that reviews with high gain-to-cost ratio cover many additional tips, while introducing little irrelevant content, and thus they should be added to the collection. Values in-between regulate the effect of efficiency in our selection.

E. Applying Greedy algorithm

Greedy algorithm is applied for making the locally optimal choice at each stage, with the hope of finding a global optimum. A greedy strategy does not in general produce an optimal solution, but nonetheless a greedy heuristic may yield locally optimal solutions that approximate a global optimal solution in a reasonable time.

F. Group of Reviews interface with Tip

In review selection from the group of reviews, micro review data consists of entities to classify or group relevant reviews. Thus seed review selection and classification is used to identify the feature selection to a group of review which matches significantly on these micro review entities. The relationship can be evaluated using the affinity between two items in the same type of entity (same dimension) or different types of entities (different dimensions) from the network. The high quality of generated reviews by the proposed algorithm will lead to efficient review selection using tip.

VI. CONCLUSION

We have extracted that micro-reviews can be used to identify the best set of reviews. We have also analyzed a set of techniques for mining and summarizing reviews. The summarized reviews are not only useful for the customers but it is also helpful for the manufacturers. The optimal solution can be derived using the Integer Linear Programming.The result which has simulated shows the better efficiencyin identifying the compact and yet informative reviews.

VII. FUTURE ENHANCEMENT

In the process of data classification, the future research will take to the accuracies,exactness and performances. It would also include methods so that this process can be evaluated in various domains. In addition to this further refinement of the summarization framework can be done.

REFERENCES

[1] Thanh-Son Nguyen, Hady W. Lauw, "*Review selection using Micro-Reviews*," in Knowledge and Data Engineering, IEEE Transaction on, 2015, No. 4, vol 27, pp. 1098-1111.

[2] K. Ganesan, C. Zhai, and E. Viegas, "*Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinions*," in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 869–878.

[3] H. Gao, J. Tang, X. Hu, and H. Liu, "*Exploring temporal effects for location recommendation on location-based social networks*," in Proc. 7th ACM Conf. Recommender Syst., 2013, pp. 93–100.

[4] E. Kouloumpis, T. Wilson, and J. Moore, "*Twitter sentiment analysis: The good the bad and the omg,"* in Proc. 5th Int. Conf. Weblogs Social Media, 2011, pp. 538–541.

[5] A. Ghose and P. G. Ipeirotis, "*Designing novel review ranking systems: Predicting the usefulness and impact of reviews*," in Proc. 9th Int. Conf. Electron. Commerce, 2007, pp. 303–310.

[6] M. Hu and B. Liu, *"Mining and summarizing customer reviews*," in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2004, pp. 168–177.

[7] P. Tsaparas, A. Ntoulas, and E. Terzi, *"Selecting a comprehensive set of reviews,"* in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2011, pp. 168–176.

[8] T. Lappas, M. Crovella, and E. Terzi, *"Selecting a characteristic setof reviews,"* in Proc. 18th ACM SIGKDD Int. Conf. Knowl.Discov.Data Mining, 2012, pp. 832–840.

