iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** II **Month of publication:** February 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue II, February 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Efficient Multi Review Classification using feature extraction technique in the Micro Reviews

Miss. A. Jenifaasheer. M.Sc, M.Phil [1], Dr.Anna Saro Vijendran. PhD[2]
[1]Research Scholar, [2]Director
Department Of Computer Science, SNR College of Arts and Science

Abstract: ***Although the content of micro-blogging sites has been studied extensively, micro-reviews are a source of content that has been largely overlooked in the literature. In this paper we study micro-reviews, and we show how they can be used for the problem of review selection. To the best of our knowledge we are the first to mine micro reviews such as foursquare tips and combine them with full-text reviews such as Yelp reviews. Our work introduces a novel formulation of review selection, where the goal is to maximize coverage while ensuring efficiency, leading to novel coverage problems. The coverage problems we consider are of broader interest, and they could find applications to different domains. We consider approximation and heuristic algorithms, and study them experimentally, demonstrating quantitatively and qualitatively the benefits of our approach. We also propose an Integer Linear Programming (ILP) formulation, and provide an optimal algorithm. This allows us to quantify the approximation quality of the greedy heuristics. Experimental results explains the performance of the system against the State of art approach in terms of precision and Recall.***
Keywords – Review Selection, Micro Reviews, Online social Networks

I. INTRODUCTION

Today data handling and management have attracted attention in research and industrial communities. Data server and Data Warehouse engineer focus on efficient management of data and its sources. Thus huge content of the reviews about company and its products is available in the web source. For instance, Yelp.com is a popular site for restaurant reviews, assisting diners to plan restaurant visits. While useful, the deluge of online reviews also poses several challenges. With the recent growth of social networking and micro blogging services [1][2], we observe the emergence of a new type of online review content. This new type of content, which we term micro-reviews, can be found in micro-blogging services that allow users to "check-in", indicating their current location or activity. For example, at Foursquare, users check in at local venues, such as restaurants, bars or coffee shops. After checking in, a user may choose to leave a message, up to 200 characters long, about their experience, effectively a micro-review of the place. Micro-reviews serve as an alternative source of content to reviews for readers interested in finding information about a place [3][4]. They have several advantages. First, due to the length restriction, micro-reviews are concise and distilled, identifying the most salient or pertinent points about the place [5]. Second, because some micro-reviews are written on site, right when the user has checked in, they are spontaneous, expressing the author's immediate and unadulterated reaction to her experience [6]. Third, because most authors check in by mobile apps, these authors are likely at the place when leaving the tips, which makes the tips more likely to be authentic. Micro-blogging sites also have the ability, if necessary, to filter out tips without an accompanying check in, thus, boosting the authenticity of the tips [7][8]. In all prior work this is modelled as a coverage problem, where the selected reviews are required to cover the different aspects of the item (e.g., product attributes), and the polarity of opinions about the item (positive and negative). To extract the aspects covered by a review and the sentiment polarity off-the-shelf tools [9] for supervised techniques are usually applied. Such approaches, although generally successful, cannot generalize to arbitrary domains. Unsupervised techniques, e.g., topic modeling [10], have also been applied (e.g., [11][12]), however they suffer from the broadness of the topic definition. We view tips as a crowd sourced way to obtain the aspects of an item that the users care about, as well as the sentiment of the users. By covering the tips, we effectively identify the review content that is important, and the aspects of the item upon which the reviews need to expand and elaborate. In our formulation, which we outline below, we make sure that the selected reviews are compact, that is, the content does not diverge from what is important about the reviewed item. We view this as an important constraint, especially for viewing on mobile devices, where screens are small, and time is short. Contributions. Although Although the content of micro-blogging sites has been studied extensively, micro-reviews is a source of content that has been largely overlooked in the literature. In this paper we study micro-reviews, and we show how they can be used for the problem of review selection. To the best of our knowledge we are the first to mine micro reviews such as Foursquare tips and combine them

with full-text reviews such as Yelp reviews. Our work introduces a novel formulation of review selection, where the goal is to maximize coverage while ensuring efficiency, leading to novel coverage problems. The coverage problems we consider are of broader interest, and they could find applications to different domains. We consider approximation and heuristic algorithms, and study them experimentally, demonstrating quantitatively and qualitatively the benefits of our approach. We also propose an Integer Linear Programming (ILP) formulation [13], and provide an optimal algorithm. This allows us to quantify the approximation quality of the greedy heuristics. We investigate the number of reviews needed to obtain perfect coverage through an alternative formulation inspired by set cover. The rest of the paper is organized as section 2 explains the background knowledge regarding the related work. Section 3 explains and formulates the proposed System. The experimental results are discussed in section 4; we conclude the work with future work of the paper at section 5.

II. RELATED WORK

A. Minimum Redundancy Feature Selection

Minimum Redundancy Maximum Relevance Feature Selection (mRMR) model is to minimize the redundancy between sequentially selected features. In the product review using a bag of words .However, this method used the greedy search, thus the global feature redundancy was considered and the results are optimal[14][15]. In this work, feature selection framework to globally minimize the feature redundancy with maximizing the given feature ranking scores, which can come from any supervised or unsupervised methods. The model has no parameter so that it is especially suitable for data mining process and review selection process in online social networks and product reviews.

B. Feature Selection using Rough Set Method

Rough set is a tool with a mathematical foundation to deal with imprecise and imperfect knowledge. It has been widely applied in machine learning, data mining and knowledge discovery. One of the applications of Rough set theory in machine learning is the so-called feature selection especially for classification problems. This is performed by means of finding a reduct set of attributes[16]. Reduct set is a subset of all features which retains classification accuracy as original attributes. Finding a reduct set in decision systems is NP-hard problem which has attracted many researchers to combine different methods with rough set.

III. OUR MODEL- REVIEW CLASSIFICATION FRAMEWORK

A. Bag of words construction from set of Micro reviews

The important sentence and the tip as bags of words are 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 can be assume that they convey a similar meaning.

B. Concept and opinion generation for Micro-reviews

Sentence and a tip may discuss the same concept (e.g., a menu dish), but use different words (e.g., soup vs. broth).In this process, it is must to determine an approximation bound for the two variants of the efficiency function in the micro review and review . Against the review selection, there exist a important aspect to determine positively and negatively opinionated sentences which are often the 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 for Set of Micro-review(Selection coverage)

Some reviews may have high coverage, but at the same time they are too verbose, containing many sentences that are not relevant to any tip at all. We would like to avoid such reviews in our selection, so we introduce the concept of efficiency. 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 2 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**.

Volume 4 Issue II, February 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

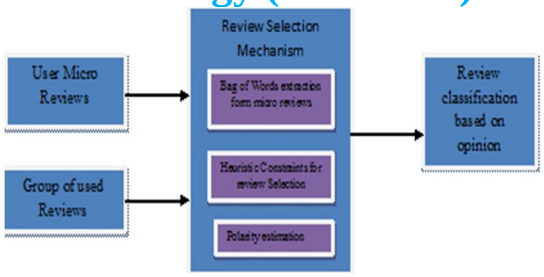


Figure 1: Review classification framework

D. Generate the set of the reviews as heuristics for Micro reviews

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. The higher the value of b, the higher the value of coverage that is needed for a low efficiency review to be included in the set.

E. Applying Greedy algorithm for selecting the review for local optimum using Selection Efficiency

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.

Greedy algorithms have five components:

1) A candidate set, from which a solution is created
2) A selection function, which chooses the best candidate to be added to the solution
3) A feasibility function, that is used to determine if a candidate can be used to contribute to a solution
4) An objective function, which assigns a value to a solution, or a partial solution, and
5) A solution function, which will indicate when we have discovered a complete solution

F. Calculate the group of Reviews interface with Tip (feature Extracted or Feature selected) using seed based review discovery (multi review selection)

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.

IV. EXPERIMENTAL RESULTS

A. Data collection and data preprocessing

The experiments require data coming from two different sources (reviews and micro-reviews), concerning the same set of entities. We pick the domain of restaurants, because it is a popular domain where there are active platforms for reviews as well as for micro-reviews. For reviews, we crawl Yelp.com to obtain the reviews of the top 110 restaurants in New York City with the highest number of reviews as of March 2012. For micro-reviews, we crawl the popular check-in site Foursquare.com to obtain the tips of the same 110 restaurants. However, some of the restaurants in Foursquare. Com have too few tips, which may not adequately reflect the restaurant's information.

Volume 4 Issue II, February 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

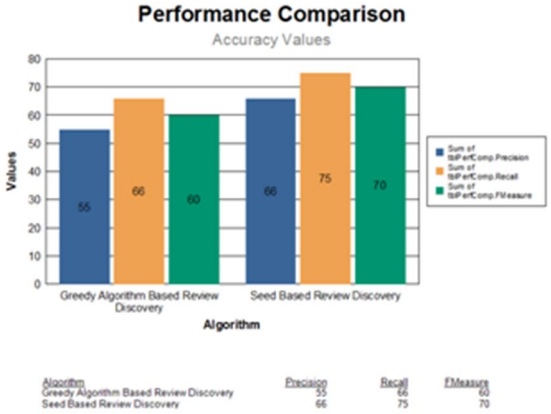


Figure 2: Performance Evaluation of feature extraction algorithm

We experiment with different values for the threshold on the probability of matching with precision and recall of the matching classifier at different values of threshold. The matching classification, and simply take all the pairs with at least one common word as matching, we get a precision of only 43%, which means more than half of all matching pairs would be incorrect.

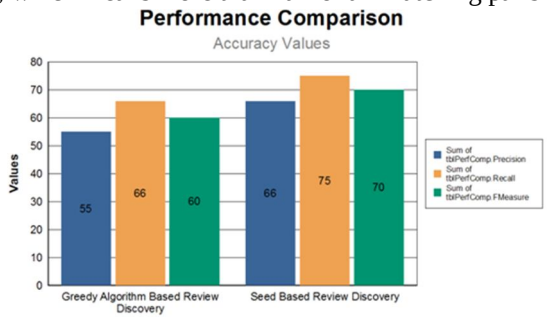


Figure 2: Performance Evaluation of feature extraction algorithm

We also experimented with temporal similarity (how close in time a tip and a review were posted) as an additional feature for the matching classifier. We found that it has essentially no effect on the accuracy, which remains practically identical.

V. CONCLUSION

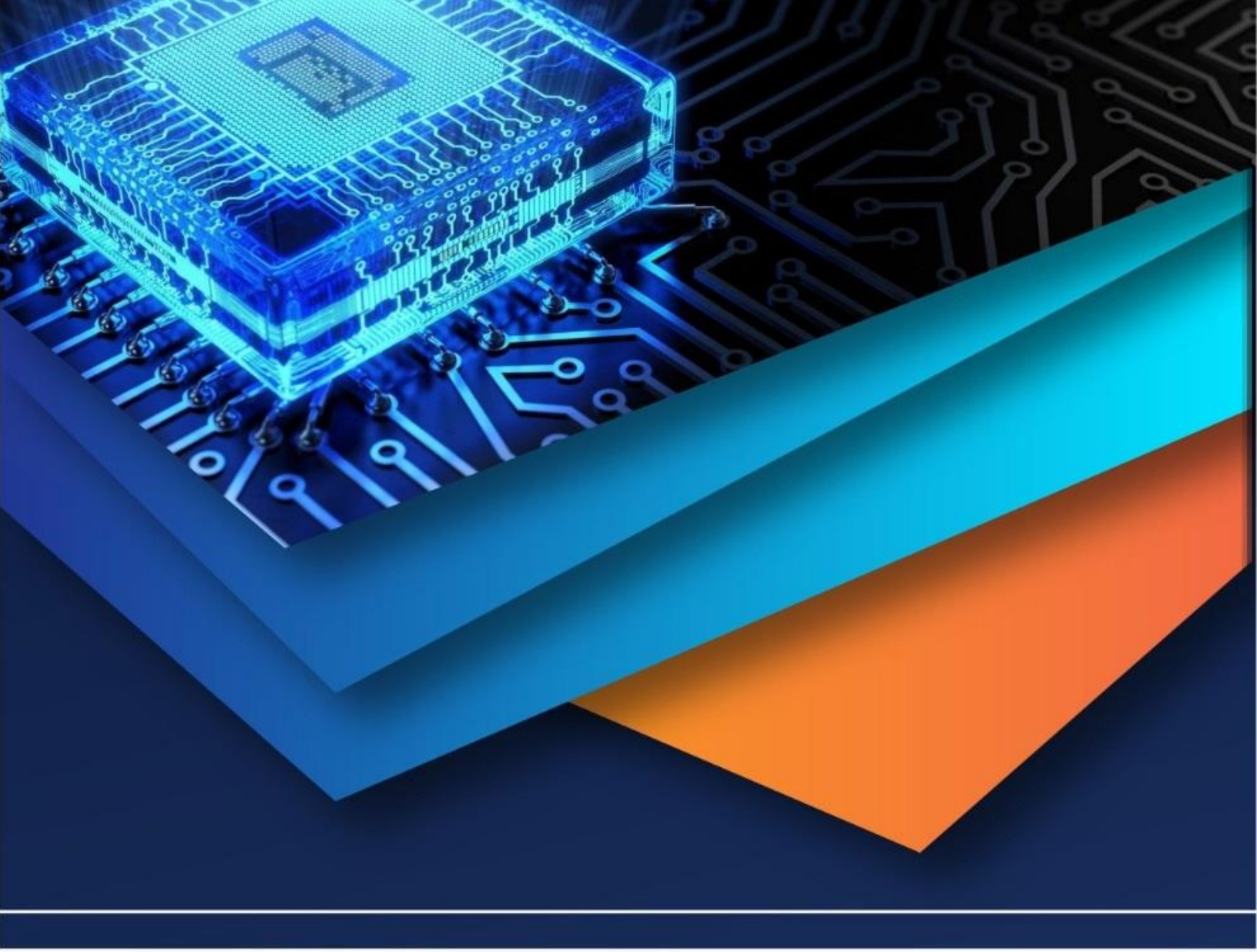
In this paper, we designed and implemented the classification technique using micro-reviews for finding an informative and efficient set of reviews. This selection criterion is for micro review is utilized to extraction of the bag of words, as well as in the efficiency constraint. The selection problem is shown to be NP-hard, and we design a heuristic algorithm EffMaxCover, which lends itself to several definitions of aggregate efficiency. The results are evaluated over a corpora of restaurants' reviews and micro reviews. Experiments show that proposed classification discovers review sets consisting of reviews that are compact, yet informative. Such reviews are highly valuable, as they lend themselves to quick viewing over mobile devices, which are increasingly the predominant way to consume Web content

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
[2] R. D. Carr, S. Doddi, G. Konjevod, and M. V. Marathe, "On the red-blue set cover problem," in Proc. 11th Annu. ACM-SIAM Symp. Discrete Algorithm, 2000, pp. 345–353.
[3] C. Cheng, H. Yang, I. King, and M. R. Lyu, "Fused matrix factorization with geographical and social influence in location based social networks," in Proc. 26th AAAI Conf. Artif. Intell., 2012, p. 1.
[4] K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions," in Proc. 23rd Int. Conf. Comput. Linguistics, 2010, pp. 340–348.
[5] K. Ganesan, C. Zhai, and E. Viegas, icropinion generation:An unsupervised approach to generating ultra-concise summaries of opinions," in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 869–878.
[6] 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.

Volume 4 Issue II, February 2016
ISSN: 2321-9653

[7] 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.
[8] 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.
[9] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," J. Amer. Soc. Inf. Sci. Technol., vol. 60, no. 11, pp. 2169–2188, 2009.
[10] S. Khuller, A. Moss, and J. S. Naor, "The budgeted maximum coverage problem," Inf. Process. Lett., vol. 70, no. 1, pp. 39–45, 1999.
[11] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in Proc. Conf. Empirical Methods Natural Lang. Process., 2006, pp. 423–430.
[12] 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.
[13] T. Lappas, M. Crovella, and E. Terzi, "Selecting a characteristic set of reviews," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov.
Data Mining, 2012, pp. 832–840.
[14] T. Lappas and D. Gunopulos, "Efficient confident search in large review corpora," in Proc. Eur. Conf. Mach. Learn. knowl. Discovery Databases: Part II, 2010, pp. 195–210.
[15] H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in Proc. Human Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2010, pp. 912–920.
[16] Y. Liu, X. Huang, A. An, and X. Yu, "Modeling and predicting the helpfulness of online reviews," in Proc. 8th Int. Conf. Data Mining, 2008, pp. 443–452.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY