Prediction of Rating from Comments based on Information Retrieval and Sentiment Analysis

Eissa M. Alshari*, Azreen Azman§, Norwati Mustapha§, Shyamala A/p C Doraisamy§ and Mostafa Alksher§

*Faculty of Computer Science
IBB University, Yemen
Email: eissa.alshari@student.upm.edu.my
§Faculty of Computer Science and Information Technology
Universiti Putra Malaysia, Serdang, Malaysia
Email: azreenazman@upm.edu.my, norwati@upm.edu.my,
shyamala@upm.edu.my, mostafa.alksher@student.upm.edu.my

***Abstract*—The number of users of an on-line shopping websites is continuously increasing. Such website often provides facility for the users to give comments and ratings to the products being sold on the websites. This information can be useful as the recommendation for other users in making their purchase decision. This paper investigates the problem of predicting rating based on users' comments. A classifier based on information retrieval model is proposed for the prediction. In addition, the effect of integrating sentiment analysis for the rating prediction is also investigated. Based on the results, an improvement in prediction performance can be expected with sentiment analysis where an increase of 54% is achieved.**

Keywords–comments; rating; information retrieval; sentiment analysis; vector space model;

I. INTRODUCTION

Due to the increasing growth in an on-line shoppings, many Internet users find it very difficult to make decision on their shopping needs. The users have to evaluate many similar products with different features, quality and prices before making a purchase decision. Unlike off-line shopping, where the customers can assess the products physically, the users are often relying on the promotional images or videos to make decision.

More recently, on-line shopping websites, such as Amazon.com, allow the Internet users to give rating for products that are being sold on the website. The rating will indicate the degree of satisfaction of the users for the particular products. In addition, those websites also provide facilities for the users to post comments regarding the products. Such comments can represent their opinion on different aspects of products. Fortunately, the ratings and comments can be used by the Internet users as additional recommendation to help them in making purchasing decisions [1].

By assuming that the ratings and comments are derived from the experience of the user's products, other users can make decision on how much that products will meet their expectation based on other's experience. However, as the number of ratings and comments for a product increases, it becomes problematic for the users to analyze the volume of the information. A summarization tool is useful to help users in analyzing the information before making decision. On the other hand, the ratings and the comments can also be used as another method to send feedback to the companies regarding their products [1].

Summarization of ratings is quite trivial, such as an average rating, can be quite informative for the users to assess the overall satisfactions on the product. Rating is often between ranges of numbers, from 1 to 5 or from 1 to 10. Therefore, users can choose a product with higher average rating for making a purchase decision. On the other hand, summarization of comments is problematic as users need to analyze the text from the comments to discover the overall opinion on the product. Sentiment analysis algorithm can be applied to the comments to extract the polarity of the text, whether are positive or negative [2][3].

Once the polarity of the texts can be discovered, the comments can be summarized based on its polarity, such as the percentage of the positive or negative. As such, summarization of ratings and comments can reduce the complexity of data analysis for the Internet users in making purchase decision. It is also noted that both ratings and comments carry different weight in decision making[4]. While rating can represent the overall satisfaction toward a product, the comment can provide much detailed description of the experience. As such, it is interesting to see how much the comments represent the rating for the product. In particular, can the rating of a product be predicted from the comment?

Many information retrieval (IR) models, such as the Vector Space Model (VSM), are used to search for relevant

978-1-5090-2954-9/16/$31.00 ©2016 IEEE

documents in response to a query by computing similarity between the text query and the text in documents [5][6]. Such a model represents the documents and the query as vectors and the similarity is estimated based on the distance between two vectors. In this context, it is hypothesize that the comments with the same ratings are more similar to each other as compared to the comments with different ratings. As such, an IR model can be used to predict the rating of product based on the comments from the users. Moreover, it is assumed that the prediction model can be improved by the integration of sentiment analysis to sense the polarity (negative or positive) of the comments.

This paper aims to investigate the problem of predicting rating of products based on comments. It is organized as follows. Related work to information retrieval and sentiment analysis techniques are discussed in Section II. The methodology for the proposed investigation is explained in Section III. In Section IV, the experimental results and analysis are presented. Finally, the conclusion is provided in Section V.

II. RELATED WORK

Many techniques have been proposed for information retrieval and sentiment analysis. A comprehensive review on text mining and sentiment analysis for unstructured web data is elaborated in [7]. The discussion in this paper suggests that the area of text mining and sentiment analysis are inter-related, where many of the techniques from data mining and natural language processing are commonly used to solve problems in this area [8].

Recently, there is an increased of interest in sentiment analysis on social media. In [9], sentiment analysis is applied to social media stream to analyze the overall sentiment of the users during disaster. During such period, people are assumed to have shown negative sentiment when discussing on the lost of lives and the damage of properties, and to have positive sentiment for inspiration and spreading hope. The investigation includes analysis of trends and geographically related sentiment based on the impact of the disaster.

In another context, trending of sentiment over the Internet and social media can give impact to the economy and financial market. Due to the ubiquity of information through search engine make it possible for the investor to create sentiment on certain issues, and may have effect to the set prices. Therefore, sentiment annotation [3] on the Internet or social media may assist users in making decision. Furthermore, attention has also been given to the implicit sentiment annotation as discussed in [10].

In many instances, sentiment analysis is applied on unstructured data from Internet which requires techniques from various related fields such as text mining, natural language processing and possibly Web crawling technologies. The basics of harnessing unstructured data from the Web and the techniques to process those data are discussed in [11], [12].

Unstructured data refers to information that doesn't have a predefined data type. Unstructured information is typically textual data, but may also contain numerical data, and factual details. This result in data that is obscure, irregular and ambiguous, thus making it difficult to analyses using conventional computing means [5].

The challenges and achievement of sentiment analysis research for social media can be observed in many evaluation tasks including those described in [13].

III. METHODOLOGY

The problem of predicting rating from comments is actually a text classification problem in which it is an attempt to classify a given text to its rating classes. In this paper, each rating is assumed to be a class and there are 5 classes (rating 1 to 5). A supervised classifier based VSM is proposed to classify text from the comments into its rating class. The classifier is then extended to use sentiment analysis to improve the classification performance.

A. Text classifier based on VSM

Vector Space Model (VSM) is one of the well-known models for IR. In retrieval, the model selects the relevant documents based on partial matching between query and all documents in the collection [14]. Both documents and the query are represented as vectors of weights for index terms and the similarity or dissimilarity between them are estimated based on the distance of two vectors. This is accomplished by assigning non-binary weights to the index terms in both query and document vectors. These terms weights are ultimately used to compute the degree of similarity between each document stored in the collection and the user query [15]. The documents are retrieved and ranked based on the degree of similarity and VSM takes into consideration those documents with partial match to the query terms.

In order to use VSM as a text classifier, a document surrogate to represent each rating class need to be constructed. A document surrogate represents the text characteristics of all comments in each rating class. Then, the similarity score is calculated between the new comment to be classified and all five document surrogate for the rating classes. The class

for the new comment is determined based on the highest similarity score. The following are the steps to construct a text classifier based on VSM.

- First, the document surrogates are constructed by aggregating all comments for each class, such that:

  $$\text{collection} = [V_i, V_i, \ldots, V_5]$$

  where Vi is the document surrogate for rating i.

- Next, in the pre-processing step, all document surrogates are tokenized and all stop words are removed. Then, stemming algorithm is applied to the remaining tokens before they are stored for indexing.

- After that, those document surrogates are indexed based on VSM model as in [16]. In short, a vector of index terms is created for each document surrogate where the weight of each term in the vector is determined by the term frequency *tf* than the inverse document frequency *idf*. In this model, a weighted term frequency w_{tf} is calculated based on [11].

  $$w_{\text{tf}} = \begin{cases} 1 + \log_{tf}, & \text{if } tf \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (1$$

  While *tf* is the frequency of the term in a document surrogate.

- Then, the inverse document frequency is calculated based on the following equation:

  $$\text{idf} = \log(1 + \frac{N}{n_t}) \quad (2)$$

  where *idf* is the inverse document frequency, N is the total number of document surrogates in the collection (i.e. 5 in this paper), and n_t is the number of document surrogates where the term t appears.

- Finally, the weight of index terms in each vector is calculated based on the following equation:

  $$\text{tf.idf} = w_{tf} \times idf \quad (3)$$

Therefore, the index will consist of five vectors representing document surrogates from each rating class. In order to determine the rating class for a new comment, a query is constructed based on the text from the new comment. A query vector is produced that consists of the weighted index terms based on the steps above. Then, similarity scores are calculated between the query and the document surrogates in the index by using cosine similarity score [17]. After all, the class for the new comment is determined based on the class for the most similar document surrogate, by using the similarity scores.

B. Extended text classifier based on VSM and sentiment analysis

It is assumed that the rating of a product has strong relationship with the sentiment of the users toward that product, based on their experience. As such, a classification model that takes into account the polarity of the comment will improve the effectiveness of the rating prediction. Therefore, this paper also aims to investigate the effect of sentiment analysis in the rating prediction.

Therefore, the comments are separated into ten classes based on the rating and the polarity (positive or negative) as opposed to five classes in the previous section. Then, a total of ten document surrogates are constructed by aggregating the comments from each class. Therefore, the collection will consist of the following:

$$\text{collection} = \begin{bmatrix} (V_1^+, V_1^-) \\ (V_2^+, V_2^-) \\ . \\ (V_5^+, V_5^-) \end{bmatrix}$$

The classifier is built based on the steps from the previous section. There are two classes for each rating, the prediction of the rating is based on the highest similarity score for the rating regardless of the polarity of the comments. For instance, if the highest similarity score belongs to the rating class 4 with positive polarity, the predicted rating class for the comment will be 4.

C. Sentiment Analysis

The classification model discussed in section III-B requires a sentiment analysis model to determine the polarity of the comment, whether it is positive or negative. In this paper, a sentiment analysis model based on sentiment lexical dictionary is used to classify comments into positive or negative labels. The lexical dictionary consists of 2005 positive terms and 4783 negative terms [18] . The approach for sentiment analysis is as follows:

- First, at the pre-processing step, all comments are tokenized into a bag of terms. What else do you do?
- Then, the number of occurrence of the positive and negative terms in the comment is calculated based on the lexical dictionary above.
- The polarity of the comment is determined based on the highest number of occurrence of the positive or negative terms in the comment.

In order to conduct the benchmark evaluation, the sentiment analysis model is evaluated by using the Large Movie Review Dataset (ACLIMDB), which is available online. The dataset consists of 25,000 highly polar movie reviews with labels [19]. The process to determine the polarity of comments based on the proposed approach

presented in [12].

The performance of the sentiment analysis is measured by positive predictive value, ***ppv***, negative predictive value, ***npv***, and the classification accuracy [17].

$$ppv = \frac{\text{tp}}{\text{tn+fp}} \quad (4)$$

Where *ppv* (precision or positive predictive value) , *tp* (true positives) and *fp* (false positive).

$$npv = \frac{\text{tn}}{\text{tn+fn}} \quad (5)$$

Where *npv* (precision or negative predictive value) , *tn* (true negative) and *fn* (false negative).

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (6)$$

The results of the sentiment analysis are summarized in Table I. The performance is described based on two approaches, with and without Porter stemmer, while the precision is measured based on [20]. Based on the table, the sentiment analysis approach with stemming is more effective than the approach without stemming. The overall accuracy of the approach is promising, with accuracy of 78.84% with stemming and 73.83% without stemming. Therefore, the sentiment analysis model for the classifier described in section III-B is implemented with stemming and stop words removal.

TABLE I
ACCURACY BETWEEN POSITIVE AND NEGATIVE PRECISION

Approaches	accuracy	positive precision	negative precision
with stem and stop words	76.84	78.76	74.93
without stem and stop words	73.83	76.26	71.41

IV. RESULTS AND DISCUSSION

In order to evaluate the effectiveness of rating prediction model proposed in this paper, a set of users comments with rating label are collected from Amazon.com. The dataset consists of 10,000 comments for training where each rating class has 2,000 comments that chosen randomly. Another 2,000 comments are selected for testing.

The experiment is conducted for two settings, the VSM classifier and the VSM classifier with sentiment analysis. For the first setting, each comment in the dataset will comprise of rating label and the textual comment with assigned ID. An example is shown in Table II.

In the second setting, polarity of each comments needs to be determined before the training of the prediction model. In this paper, the sentiment analysis method described in Section III-C is used to classify each textual comments to its polarity, whether positive of negative. The method has been benchmarked by using (ACLIMDB) dataset, and the performance has been elaborated in Section III-C. Therefore, now each comment will comprise of rating label, the textual comment and its polarity with assigned ID, as shown in Table III.

TABLE II
SAMPLE OF AMAZON DATASET

ID	Rate	Text
981670	5	*In my review I meant to mention that theres a good book out right now about James Taylor So far its a good read I bought it here Fire and Rain by Ian Halperin Check it out*

TABLE III
SAMPLE OF AMAZON DATASET WITH SENTIMENT POLARITY

ID	Rate	Polarity	Text
981670	5	positive	*In my review I meant to mention that theres a good book out right now about James Taylor So far its a good read I bought it here Fire and Rain by Ian Halperin Check it out*

In addition, two weighting schemes are used for the experiment, namely the *idf* only and *tf.idf* . As shown in Table IV, the average prediction precision for VSM classifier is 32.1% for *idf* weighting scheme and 38.2% for *tf.idf* weighting scheme, which shows that *tf.idf* it is a better weighting scheme for VSM classifier. However, the opposite pattern is observed for VSM classifier with sentiment analysis where the *idf* weighting schemes is better with 49.5% as compared to *tf.idf* weighting scheme with 38.9% prediction precision.

TABLE IV
PREDICTION PRECISION OF VSM CLASSIFIER WITH SENTIMENT ANALYSIS

Prcesion	idf	tf.idf
VSM_classifier	32.1	38.2
VSM_classifier_with sentiment	49.5	38.9

A positive impact can observed for integrating sentiment analysis with VSM classifier, such that the prediction performance is improved. In the case of *tf.idf* weighting scheme, there is only small improvement, which is about 2%. On the other hand, a significant improvement can be observed for *idf* weighting scheme with 54% increase in prediction precision. Figure 1 shows the overall performance of the prediction models.

V. CONCLUSION

This paper investigates the problem of predicting rating based on the comments from the Internet users. A classifier

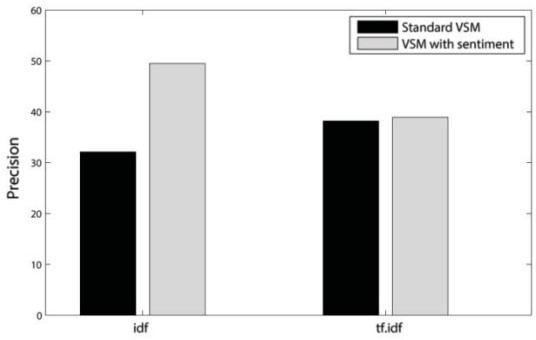

Figure 1. The performance of the prediction models

based on information retrieval and the Vector Space Model are proposed to solve the problem. The effect of integrating sentiment analysis model into the classifier is also investigated. The outcome promising showing that sentiment analysis has positive effect to the performance of the classifier in predicting rating form textual comments of the users.

In future, more models of information retrieval, such as the probabilistic model or the statistical language model, can be adopted as the classifier to improve the prediction performance. In addition, the sentiment analysis model can be improved with advanced NLP techniques.

REFERENCES

[1] J. B. Schafer, J. A. Konstan, and J. Riedl, "E-commerce recommendation applications," in *Applications of Data Mining to Electronic Commerce*. Springer, 2001, pp. 115–153.

[2] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," *Information Sciences*, vol. 311, pp. 18–38, 2015.

[3] Z. Da, J. Engelberg, and P. Gao, "The sum of all fears investor sentiment and asset prices," *Review of Financial Studies*, vol. 28, no. 1, pp. 1–32, 2015.

[4] B. Agarwal, N. Mittal, P. Bansal, and S. Garg, "Sentiment Analysis Using Common-Sense and Context Information," *Computational Intelligence and Neuroscience*, vol. 2015, pp. 1–9, 2015. [Online]. Available: http://www.hindawi.com/journals/cin/2015/715730/

[5] K.-y. Chen, H.-s. Lee, H.-m. Wang, B. Chen, and H.-h. Chen, "I-Vector Based Language Modeling for Spoken Document Retrieval," pp. 7133–7137, 2014.

[6] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1–12, 2013. [Online]. Available: http://arxiv.org/pdf/1301.3781v3.pdf

[7] R. Nikhil, N. Tikoo, S. Kurle, H. S. Pisupati, and G. Prasad, "A survey on text mining and sentiment analysis for unstructured web data," in *Journal of Emerging Technologies and Innovative Research*, vol. 2, no. 4 (April-2015). JETIR, 2015.

[8] C. C. Aggarwal and C. Zhai, *A Survey of Text Clustering Algorithms*, 2012.

[9] Y. Lu, X. Hu, F. Wang, S. Kumar, H. Liu, and R. Maciejewski, "Visualizing social media sentiment in disaster scenarios," in *Proceedings of the 24th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2015, pp. 1211–1215.

[10] M. Van de Kauter, B. Desmet, and V. Hoste, "The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment," *Language Resources and Evaluation*, pp. 1–36, 2015.

[11] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of twitter," *Information Processing & Management*, 2015.

[12] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[13] S. Rosenthal, P. Nakov, S. Kiritchenko, S. M. Mohammad, A. Ritter, and V. Stoyanov, "Semeval-2015 task 10: Sentiment analysis in twitter," in *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval*, 2015.

[14] E. Emad, E. M. Alshari, and H. Abdulkader, "Boolean information retrieval based on semantic," in *International conference on intelligent computing and information*, vol. 8, no. 6. Ain Shams, 2013, pp. 87–93.

[15] S. Bergsma and Q. I. Wang, "Learning Noun Phrase Query Segmentation," *Computational Linguistics*, no. June, pp. 819–826, 2007.

[16] E. M. Alshari, "Semantic arabic information retrieval framework," Ph.D. dissertation, Menufia University, Egypt, 2014.

[17] E. Emad, E. M. Alshari, and H. Abdulkader, "Arabic vector space model based on semantic," in *International journal of computer science (IJISI),*, vol. 8, no. 6. Ain Shams, 2013, pp. 94–101.

[18] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[19] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150. [Online]. Available: http://www.aclweb.org/anthology/P11-1015

[20] P. K. Singh and M. S. Husain, "Methodological study of opinion mining and sentiment analysis techniques," *International Journal on Soft Computing*, vol. 5, no. 1, p. 11, 2014.