A Framework for Automated Rating of Online Reviews against the Underlying Topics

Xiangfeng Dai, Irena Spasić, Frédéric Andrès

ABSTRACT: Even though most online review systems offer star rating in addition to free text reviews, this only applies to the overall review. However, different users may have different preferences in relation to different aspects of a product or a service and may struggle to extract relevant information from a massive amount of consumer reviews available online. In this paper, we present a framework for extracting prevalent topics from online reviews and automatically rating them on a 5-star scale. It consists of five modules, including linguistic pre-processing, topic modelling, text classification, sentiment analysis, and rating. Topic modelling is used to extract prevalent topics, which are then used to classify individual sentences against these topics. A state-of-the-art word embedding method is used to measure the sentiment of each sentence. The two types of information associated with each sentence – its topic and sentiment – are combined to aggregate the sentiment associated with each topic. The overall topic sentiment is then projected onto the 5-star rating scale. We use a dataset of Airbnb online reviews to demonstrate a proof of concept. The proposed framework is simple and fully unsupervised. It is also domain independent, and, therefore, applicable to any other domains of products and services.

KEYWORDS: natural language processing, topic modelling, machine learning, visualization, sentiment analysis, latent dirichlet allocation, weighted word embeddings, data mining, big data

1 INTRODUCTION

Online reviews are valuable sources of relevant information that can support users in their decision making. An estimated 92% of online shoppers read online reviews, 88% trust online reviews as much as personal recommendations and they typically read more than 10 reviews to form an opinion [1]. The objective of this study is to propose a framework aimed at improving user experience when faced with an otherwise unmanageable amount of online reviews. This is achieved by automatically extracting the underlying topics (e.g. a review of a garment of clothing may contain different opinions about different aspects of the product such as fit, fabric, color or pattern, craftsmanship, etc.) and rating reviews with respect to these topics. The rating framework combines algorithms for topic modelling, text classification, and sentiment analysis.

Most approaches to rating of online reviews use supervised learning approaches. For instance, Hu and Liu [2] manually annotated 2,006 positive words and 4,783 negative words to train classifiers used to analyze customer reviews. Similarly, Ganu et al. [9] rated 52,264 restaurant reviews after manually annotating a set 3,400 sentences with category and sentiment labels in order to train an SVM classifier. However, the process of manually annotating large training datasets is labour- and time-intensive. Moreover, such approaches are not readily portable to other domains.

A variety of studies on rating online reviews have been published for a wide range of domains. Chevalier et al. [8] predicted ratings of online book reviews. Dellarocas and Zhang [10] demonstrated a case of rating movie reviews. However, these studies did not take into account the sentiment or semantic content of text reviews. Furthermore, the studies [9] [10] [11] [19] [25] [26] [27] [28] [29] focused on predicting ratings in a specific domain such as: restaurants, tourism, movies,

hotels, healthcare, etc. These approaches are domain-dependent and cannot be easily implemented and transferred to other products and services.

2 CHALLENGES

- **Large Volume**: The large volume of online reviews creates significant information overload [2] [9] [10] [16] [17] [18] [20] [21]. It is challenging to uncover underlying topics from a massive amount of online reviews and especially to rate them against these topics.
- **Informality**: Online reviews are informal documents in terms of style and structure [2] [3] [9] [12] [18] [21]. The language used may contain abbreviations, slang, spelling mistakes, typographical errors, special characters, hyperlinks, redundant whitespaces, etc.
- **Supervision:** Sentiment analysis plays an important role in predicting ratings from text reviews [2] [9] [12] [18]. Supervised and semi-supervised classification methods require a large amount of manually annotated instances to train a sentiment classifier. The manual annotation process is time- and labour-intensive.
- **Context-awareness**: The vast majority of sentiment classification approaches rely on the bag-of-words model, which disregards context, grammar and even word order [3]. Approaches that analyse the sentiment based on how words compose the meaning of longer phrases have shown better result [31], but they incur an additional annotation overhead.
- **Domain independence:** Any implementation should ideally be portable from one domain to another. In particular, the performance should not depend significantly on lexical resources that need to be hand-crafted for a particular domain [18] [19].

3 FRAMEWORK DESIGN AND METHODOLOGY

We present a framework for rating online reviews, which extracts the underlying topics automatically and rates each review against these topics. The framework consists of five modules, including linguistic pre-processing, topic modeling, text classification, sentiment analysis and rating (Fig. 1). The following subsections provide details about each module.

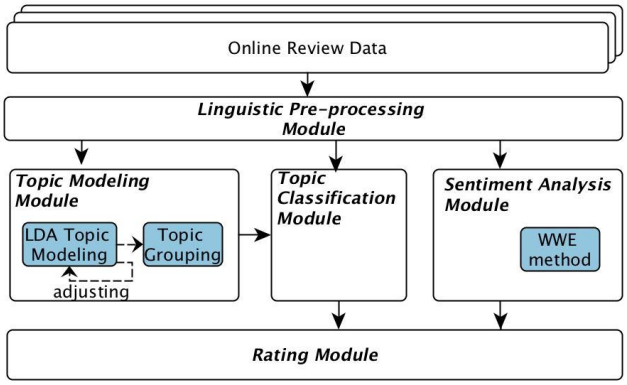

Figure 1: Framework for Rating Online Reviews

3.1 Linguistic Pre-processing

We have previously discussed the challenges associated with automated analysis of online reviews, including the lack of formal structure and informal style of writing. To prepare the raw text for further analysis, including topic modelling and sentiment analysis, we employed the following linguistic pre-processing steps [3] [12]:

- Removing stop words.
- Correcting spelling mistakes and typographical errors.
- Converting slang and abbreviations to the corresponding words.
- Stemming to aggregate words with related meaning.
- Tokenization.
- Removing punctuation, special characters, hyperlinks, etc.

3.2 Topic Modelling

In this module, we use Latent Dirichlet Allocation (LDA), an unsupervised probabilistic method that is widely used to automatically discover underlying topics from a set of text documents based on word distribution [6][7]. To demonstrate the approach, we used a publicly available dataset of Airbnb online reviews [13]. Each topic is represented as a collection of words with Dirichlet distribution [7]. Each review may be associated with multiple topics. Table 1 shows three examples of topics represented by 10 most relevant words within a topic. Intuitively, according to the given words, one may assume that the topic T1 is related to amenities, whereas T2 and T3 are more about the location.

Table 1: Example of LDA topic modeling

ID	LDA Models
T1	0.033*bathroom + 0.027*floor + 0.023*unit + 0.021*door + 0.017*building + 0.016*fine + 0.016*location + 0.015*bedroom + 0.014*air + 0.014*people
T2	0.057*station + 0.055*walk + 0.048*arrival + 0.044*house + 0.043*day + 0.036*subway + 0.031*center + 0.026*city + 0.026*train + 0.023*neighborhood
T3	0.133*location + 0.036*distance + 0.035*joe + 0.026*walk + 0.022*check + 0.020*neighborhood + 0.018*city + 0.017*close + 0.016*airport + 0.015*convenient

The number of topics is an input parameter to the LDA method, which is related to their coverage and their comprehensibility. In a series of experiments and manual inspection of the generated topics, we decided to restrict the number of topics to 10 and the number of feature words to 3000 most frequent ones [30]. Some automatically extracted topics might be similar (e.g., both T2 and T3 are related to the location aspect), and we aggregated such topics manually. Overall, we arranged 10 topics into four themes: Location, Amenities, Family-friendliness and Other.

3.3 Topic Classification

Once the topic model has been generated, each sentence can be checked against the model to obtain information on topic distribution, which can be used to classify the sentence into an appropriate topic [6][7] (see Table 2 for examples).

Table 2: Examples of Topic Classification

Sentence	Classification
This spot has a perfect location as it's nestled in a quiet neighborhood yet only a 3 min walk to the St Mary's Green Line C stop.	Location
Dirty grubby apartment with single AC in one bedroom so rest of apartment was boiling, broken TV Screen, kitchen smelt of Gas, broken closet door, cracked tiles in bathroom that smelt damp, filthy stairs carpet, and dirty paint work.	Amenities
The kitchenette was perfect for medium cooking as it has a stove, refrigerator and has cleaning supplies to cleanup after!	Amenities

3.4 Sentiment Analysis

We operate under an assumption that the rating is correlated with the sentiment strength. To calculate the overall sentiment, each sentence is analyzed separately using the weighted word embeddings method [3] [5]. The word embedding algorithm can capture semantic relationships from the surrounding words and has the advantage of being unsupervised, i.e. not requiring manual annotation of a large training dataset. Once all sentences have been analyzed, the sentiment associated with each topic is aggregated across the relevant sentences.The following steps provide more detail about our sentiment analysis approach.

Step 1: The sentiment score of each word represented by a vector is calculated based on the cosine similarity between its vector of a word and the vectors of seed words of positive and negative sentiments as they are defined in [3].

$$pws(w) = sim(seed_{pos}, w) - sim(seed_{neg}, w) \qquad (1)$$

$$sim(A, B) = \frac{\sum_{i=1}^{n}(A_i * B_i)}{\sqrt{\sum_{i=1}^{n} A_i^2 * \sum_{i=1}^{n} B_i^2}} \qquad (2)$$

where *A* and *B* are vectors of length *n*.

Step 2: Negation Handling – Negation words and punctuation marks are used to determine the context affected by negation. We predefined a list of negation words such as “no” or “not”. If a negation word appears within a predefined distance (e.g. one token before and two tokens after the negation word), the sentiment polarity of words within the negated context is inverted.

Step 3: Part-of-Speech Tagging – Not every word is equally important for sentiment analysis, e.g. most sentiment words are adjectives, adverbs, nouns and verbs [32]. To automatically identify such words, we use NLTK toolkit for parts-of-speech tagging [22].

Step 4: Having calculated the sentiment of individual words as described in Step 1, the sentiment of a sentence is calculated using the following formula:

$$\text{PSS} = \sum_{j=1}^{K} weight(j) * pws(j) \qquad (3)$$

where K is the total number words in the sentence, $weight(j)$ is the part-of-speech weight of the j^{th} word and $pws(j)$ is the sentiment score of the j^{th} word.

Over 2M experiments using 256 combinations of part-of-speech weights were previously conducted [3], based on which 9 combinations were recommended for the weighted word embeddings for sentiment classification. In this study, we used the weights 1, 3, 2, and 2 for nouns, verbs, adjectives, and adverbs respectively. Table 3 shows the examples of automatically scored sentiment for the given sentences. The sentiment score indicates the polarity of the sentence: the first and third sentences are positive, the second sentence is negative. The sentiment score also reflects the strength of the overall sentiment, e.g. the first sentence and the third sentence are both positive, but the sentiment of the first sentence is stronger than that of the third sentence.

Table 3: Examples of Automatically Computed Sentiment Score

Sentence	PSS
This spot has a perfect location as it's nestled in a quiet neighborhood yet only a 3 min walk to the St Mary's Green Line C stop.	3.46
Dirty grubby apartment with single AC in one bedroom so rest of apartment was boiling, broken TV Screen, kitchen smelt of Gas, broken closet door, cracked tiles in bathroom that smelt damp, filthy stairs carpet, and dirty paint work.	-0.58
The kitchenette was perfect for medium cooking as it has a stove, refrigerator and has cleaning supplies to cleanup after!	1.72

3.5 Topic Rating

Once the topic model has been extracted from a corpus of reviews, each sentence is classified into an appropriate topic. To rate a whole review against the given topics, we used the sentiment of all sentences associated with each topic. To rate a review from on a 5-star scale (1 star being very negative and 5 star being very positive), we first normalize the sentiment score of each sentence as follows:

$$\text{score} = 5\frac{\text{PSS} - \text{PSS}_{\text{min}}}{\text{PSS}_{\text{max}} - \text{PSS}_{\text{min}}} \qquad (4)$$

where PSS_{min} and PSS_{max} are the minimum and maximum sentiment score in a text review, and PSSis the sentiment score of the given sentence. The normalization effectively maps the sentiment of each sentence to a real number between 0 and 5.

For each topic in turn, we aggregate the normalized scores of all sentences within the topic to obtain the average score, denoted as $\text{score}_{\text{ave}}$. We then map the average score to 5-star rating using the rules given in Table 4.

Table 4: Star Ratings

Score	Stars
$0 < \text{score}_{ave} \leq 1$	★
$1 < \text{score}_{ave} \leq 2$	★★
$2 < \text{score}_{ave} \leq 3$	★★★
$3 < \text{score}_{ave} \leq 4$	★★★★
$4 < \text{score}_{ave} \leq 5$	★★★★★

4 EXPERIMENTS

4.1 Data

Airbnb Online Review Dataset: We used the Boston Airbnb Open Data, a publicly available set of reviews [13]. The dataset contains 68,276 user reviews of 3,586 Airbnb listings. As part of the Airbnb Inside initiative, this dataset describes the listing activity of home stays in Boston, MA.

Google News Dataset (Word2vec Model): Google's pre-trained vector set [4] is used in the sentiment analysis module. The model contains 300-dimensional vectors for 3 million words and phrases.

4.2 Implementation and Results

The core algorithms were implemented with Python, NLTK [22] toolkit and Gensim [23] library. All reviews were stored in MongoDB [24] for easy access and processing. It uses JSON-like documents with schemas and is useful for very large sets of distributed data. The front-end pages used to visualize the results were developed with HTML5, JavaScript and CSS.

Fig. 2 provides an example of topic-related ratings for a given review. Different topics are highlighted in different colors. Each sentence is tagged with its sentiment score and topic classification at the end. The overall ratings of the given review in terms of location and amenities were calculated as 4-stars and 3-stars respectively.

Figure 2: Example of topic-specific ratings

5 CONCLUSIONS

In this study, we presented a framework for rating online reviews against automatically extracted underlying topics. The proposed framework consists of modules: (1) linguistic pre-processing, (2) topic modeling, (3) sentence classification against the topics extracted in the previous module, (4) sentiment analysis, (5) rating against the topics based on the sentiment of the corresponding sentences. The proposed method is unsupervised, i.e. does not require an annotated training dataset. It is also domain independent, and, therefore, can be applied across different domains for which online reviews are available. As part of future work, we will formally evaluate the effectiveness of this method on a variety of domains and datasets.

REFERENCES

[1] K. Shrestha, "50 Stats You Need to Know About Online Reviews", https://www.vendasta.com/blog/50-stats-you-need-to-know-about-online-reviews, (Accessed: 2 Mar 2017)

[2] Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." *In Proceedings of the tenth ACM SIGKDD international conference* ACM, 2004.

[3] X. Dai and B. Prout. 2016. Unlock big data emotions: Weighted word embeddings for sentiment classification. *In Big Data (Big Data), 2016 IEEE International Conference. IEEE*.

[4] Google word2vec. https://code.google.com/archive/p/word2vec/ (Accessed: 24 Feb. 2017)

[5] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space*", ICLRWorkshop*, 2013.

[6] Shu, Liangcai, Bo Long, and Weiyi Meng. "A latent topic model for complete entity resolution." Data Engineering, *IEEE*, 2009.

[7] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* , 2003

[8] Chevalier, Judith A., and Dina Mayzlin. "The effect of word of mouth on sales: Online book reviews." *Journal of marketing research* 43, no. 3, 2006

[9] G. Ganu, N. Elhadad, and A. Marian. Beyond the stars: Improving rating predictions using review text content. *In WebDB, volume 9, pages* 1{6, 2009.

[10] Dellarocas, Chrysanthos, Xiaoquan Michael Zhang, and Neveen F. Awad. "Exploring the value of online product reviews in forecasting sales: The case of motion pictures." *Journal of Interactive marketing* 21, no. 4, 2007.

[11] Ye, Qiang, Ziqiong Zhang, and Rob Law. "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches." *Expert Systems with Applications* 36, no. 3, 2009.

[12] X. Dai and R. Prout. 2016. Unlocking Super Bowl Insights: Weighted Word Embeddings for Twitter Sentiment Classification. *In Proceedings of the 3rd Multidisciplinary International Social Networks Conference on SocialInformatics* 2016, *Data Science* 2016, ACM.

[13] Airbnb, "Boston Airbnb Open Data", https://www.kaggle.com/airbnb/boston, (Accessed: 12 February 2017)

[14] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Multi-facet rating of product reviews. *In Advances in Information Retrieval, 31th European Conference on Information Retrieval Research (ECIR 2009*), 2009.

[15] M. P. O'Mahony, P. Cunningham, and B. Smyth. An assessment of machine learning techniques for review recommendation. *In Proceedings of the 20th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2009),* pages 244–253, Dublin, Ireland, 2009.

[16] A., Rajeev, A. Kadadi, X. Dai, and F. Andres. "Challenges and opportunities with big data visualization." In Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems, pp. 169-173. ACM, 2015.

[17] Ganu, Gayatree, Yogesh Kakodkar, and AméLie Marian. "Improving the quality of predictions using textual information in online user reviews." *Information Systems* 38, no. 1 (2013): 1-15.

[18] B Samuel and N Elhadad. "An unsupervised aspect-sentiment model for online reviews." *Association for Computational Linguistics*, 2010.

[19] P. Alexis and F. Knolle. "Exploring the adoption and processing of online holiday reviews: A grounded theory approach." *Tourism Management* 32, no. 2 (2011): 215-224.

[20] G. Anindya and P Ipeirotis. "Designing novel review ranking systems: predicting the usefulness and impact of reviews." *In Proceedings of the ninth international conference on Electronic commerce*, ACM, 2007.

[21] G. Anindya, and P. Ipeirotis. "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics." *IEEE Transactions on Knowledge and Data Engineering*, 2011.

[22] Natural Language Toolkit. http://www.nltk.org/ (Accessed: 28 Feb. 2017)

[23] Gensim. https://radimrehurek.com/gensim/about.html (Accessed: 26 Feb. 2017)

[24] MongoDB. https://www.mongodb.com/ (Accessed: 15 Feb. 2017)

[25] Joshi, Mahesh, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. "Movie reviews and revenues: An experiment in text regression." *In Human Language Technologie*s, 2010.

[26] Zhuang, Li, Feng Jing, and Xiao-Yan Zhu. "Movie review mining and summarization." In Proceedings of the 15th ACM international conference. ACM, 2006.

[27] Fang, Bin, Qiang Ye, Deniz Kucukusta, and Rob Law. "Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics." Tourism Management 52, 2016

[28] H. Lee and S. Blum. How hotel responses to online reviews differ by hotel rating: an exploratory study. *Worldwide Hospitality and Tourism Themes*, 2015.

[29] X. Ruan. "A five-star doctor? Online rating of physicians by patients in an internet driven world." *Pain physician* 18, 2015

[30] C. Manning, P. Raghavan and H. Schütze, An Introduction to Information Retrieval, Online edition, Cambridge University Press, 2009.

[31] Socher et al. "Recursive deep models for semantic compositionality over a sentiment treebank." In Proceedings of the conference on empirical methods in natural language processing (EMNLP), 2013.

[32] Liu, Bing. "Sentiment Analysis and Opinion Mining Morgan & Claypool Publishers." Language Arts & Disciplines, 2012