Master paper outline & draft

Title: Observe the sentiment patterns of Airbnb customer reviews by aspect scores based on sentence-aspect relevance model and NLP knowledge

Does reviews of Airbnb home correspond to its aspect scores of Airbnb? Evidence from sentiment analysis of reviews using proposed sentence-to-aspect relevance model.

## Introduction

**Paragraph 1:** Motivations to study Airbnb customer reviews from the sentiment analysis perspective

## Literature review

**Part 1: previous studies on sentiment analysis**

Sentiment analysis is an area that has cumulative researches in a lone time. There are many approaches to this field in which the two commonly studied approaches are the corpus-based approach and the lexicon-based approach(Liu, B. 2012). Each approach has its pros and cons. In the paper (Abdulla, Mahyoub, Shehab, & Al-Ayyoub, 2013) and (Itani, Zantout, Hamandi, & Elkabani, 2012) compared the two approaches while another paper (El-Halees et al., 2011) proposed to merge them into hybrid approaches.

In general, sentiment analysis has been investigated mainly at three levels:

1. Document-level: classify what opinion does a whole document expresses, positive or negative sentiment (Pang, Lee, and Vaithyanathan, 2002). This kind of analysis assumes that each document expresses opinions on a single entity.

2. Sentence level: investigating positive, negative, or neutral opinion expressed by each sentence. This level of analysis is closely related to subjectivity classification (Wiebe, Bruce, and O'Hara, 1999), which distinguishes sentences that express factual information from sentences that express subjective views and opinions.

3. Entity and Aspect level: Aspect level also called feature level performs more detailed analysis. Instead of looking at documents, paragraphs, sentences, aspect level directly looks at the opinion itself. It is based on the idea that opinion consists of sentiment (positive or negative) and a target (of opinion). An opinion without its target being identified is of limited use. Based on this level of analysis, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyses.

**Part 2: Aspect-based Sentiment Analysis**

Reviews conveying different sentiments about different aspects of the same product require an aspect-based sentiment analysis approach to process them (Liu, 2012; Piryani, Madhavi, & Singh, 2016). Compared with traditional sentiment analysis, aspect-based sentiment analysis need to face some additional challenges such as: linking each part of the text to the aspect refers to; identifying the parts of text discussing the same aspect and dealing with comparative sentences, etc. (Cambria, Schuller, Xia, & Havasi, 2013).

As for the supervised learning part, in (Wei and Gulla, 2010), a hierarchical classification model was also proposed with the key issue as of how to determine the scope of each sentiment expression such as whether it covers the aspect of interest in the sentence. Currently, one main approach is to use parsing to determine the dependency and the other relevant information. For example, a related approach was also used in (Boiy and Moens, 2009), which weights each feature based on the position of the feature relative to the target aspect in the parse tree. However, the methods are all mainly used for document-level sentiment classification because of enough length of documents which allows doing features extraction for classification than individual sentences like reviews. Thus, supervised learning has difficulty in dealing with online review analysis.

Then, another kind of learning, unsupervised learning, found some possible and efficient ways. The lexicon-based approach can avoid some of the issues (Ding, Liu, and Yu, 2008; Hu and Liu, 2004), and has been shown to perform quite well in a large number of domains. They use a list of sentiment words, phrases, expressions, rules of opinions and so forth, to determine the sentiment polarity on each aspect in a sentence.

**Part 3: Previous researches of Airbnb reviews or online product reviews**

* The characteristics of Airbnb reviews that need to pay attention to or may affect this research.
  + Who write the online reviews
  + The sentiment bias of Airbnb reviews
* Experiences from other similar online review sentiment analysis case such as hotel reviews, etc.

**Part 4: The relevance between sentence and aspect**

* Word-to-word similarity model (Word2Vec)
* Bag-of-words comparison (term frequency by weight)

## Methodology

**Part 1: Data selection process and reason**

**1.1 Initial dataset: Airbnb officially released open data on Kaggle**

The initial dataset will be the official open data released by Airbnb about Seattle, WA. Detailed information of the dataset can be seen in the following bullet points.

The tables and columns used in this research are “Listings” and “Reviews” tables which contain the required data for finding the aspect-relevant sentences in review texts, using sentiment detection model based on aspects: the ids, prices, average scores of all 6 aspects, review texts of Airbnb homes.

* Source: <https://www.kaggle.com/airbnb/seattle>
* This dataset describes the listing activity of homestays in Seattle, WA.
* The following Airbnb activity is included in this Seattle dataset:
  + Listings, including full descriptions and average review score
  + Reviews, including unique id for each reviewer and detailed comments
  + Calendar, including listing id and the price and availability for that day
* Seattle dataset uses the following parameters (from Airbnb Inside):
  + A high availability metric and filter of 60 days per year to help identify listings that might be available permanently, rather than occasionally
  + A frequently rented filter of 60 days per year to identify properties that are being used full-time as a hotel, not occasionally, and taking away residential supply
  + A review rate of 50% for the number of guests making a booking who leave a review
  + An average booking of 3 nights unless a higher minimum night is configured for a listing.
  + A maximum occupancy rate of 70% to ensure the occupancy model does not produce artificially high results based on the available data

**Reasons:**

* The authenticity and convincingness of data source
  + The released official dataset has preliminary parameters to avoid occasionality of Airbnb homes and then reduce the risks of scraping data by researcher herself such as having a large proportion of outlier data in scraped data. For example, data obtained by own-written spider may be relevant to those shortly opened homes due to the poor ability in distinguishing their usage period.
* Well-structured CSV files are applicable to most of the data analysis tools and are easy to be loaded using programming language. This could save this master paper research a considerable amount of time.

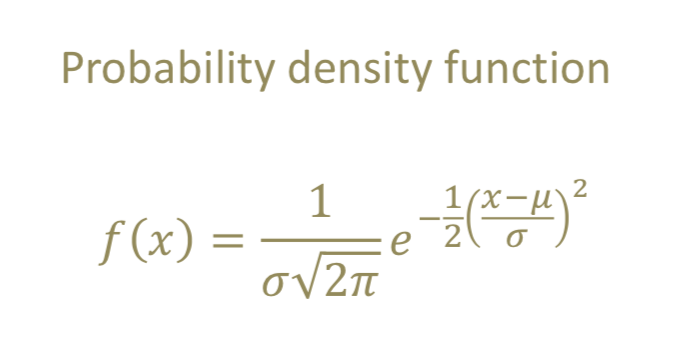
**1.2 Data clean and preliminary selection: remove unused data**

**1.2.1 Data cleaning**

* Remove N/A data.
  + removed 660 from 3818 homes which are lack of some score information
* Select necessary columns
  + Selected 11 columns in “listings.csv” and 3 columns in “reviews.csv”
* Change data types to facilitate further analysis

**1.2.2 Preliminary Data Selection**

This research aims to analyze the majority of all homes in Seattle but still need to remove some homes that in the tail of all homes’ distribution. The “Price” attribute is the chosen factor to draw the distribution of all homes. The research will use the knowledge of Gaussian Distribution also known as Normal Distribution.

* Train all Airbnb homes to Gaussian Distribution on price
  + The confidential interval in distribution can guarantee the sample data selection this research are the majority of all homes.
* PDF (probability density function) and CDF (Cumulative distribution function) concept will be used.
* To guarantee the probability of homes with prices within the selected price range has is 80%

**Part 2: Sentence to aspect detection model by weighted term-frequency**

**2.1 The selection of words related to aspects**

First step: Keywords from the Airbnb official text description of each aspect

* For example:
  + Location: How did guests feel about your neighborhood?
  + Keyword: location, neighborhood

Second step: Use WordNet, a large lexical database of English, to extract the relevant bag-of-words of each aspect. The words should contain both synonyms and antonyms. The detailed process is:

* Look through the description of synsets, for example, the “location” aspect:
  + location .n.01: a point or extent in space
  + placement .n.03: the act of putting something in a certain place
  + localization .n.01: a determination of the place where something is
  + location .n.04: a workplace away from a studio at which some or all of a movie may be made
* Exclude irrelevant synset
  + From above, we can see the location.n.04, localization.n.01, placement.n.03 are not necessarily in accord to the Airbnb's definition of aspect location. So, I will ignore synonyms in those synsets.
* Retrieve the antonym synsets for relevant synsets.
* Consist of the bag-of-words of aspects by getting synonyms from relevant synsets and antonyms from antonym synsets.

Last step: build a dictionary, aspect\_keywords\_dic, which contains aspects and their relevant keywords.

**2.2 Find the relevance between sentences in review texts and scoring aspects.**

**2.2.1 Processing of sentences in review texts**

Using spaCy, an open-source software library for Natural Language Processing, to segment sentences from review paragraphs, compute the similarity between words in aspect bag-of-words and words in review sentences.

**2.2.2 Map sentences to its relevant aspect**

For each sentence, compute an aspect-similarity-score which starts from 0 and will add one once a word in a sentence has the word-similarity-score larger than a relevance score(such as 0.5) with any word in the aspect’s bag-of-words.

The similarity score between two words, word-similarity-score, will be computed using Word2Vec similarity model. The model is to maximize the log of the dot product of the word vector and its context word vector. Higher the magnitude of the dot product, higher would be the cosine similarity.

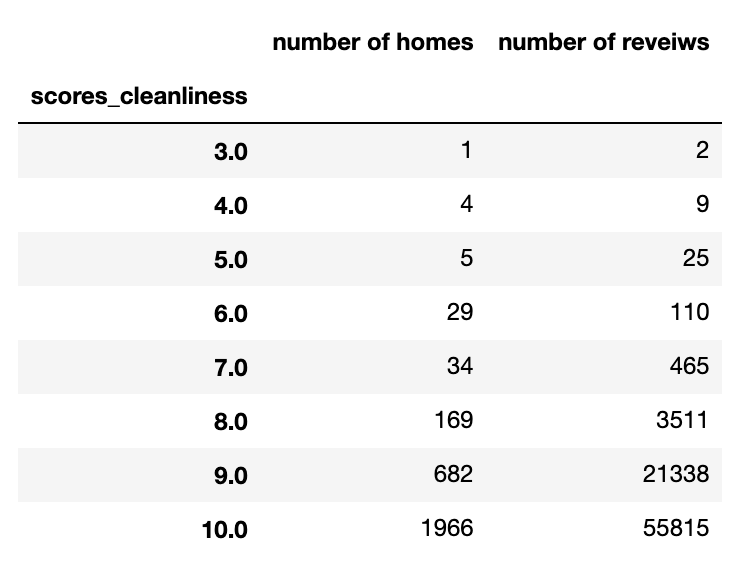
Adding weight to words in a sentence by the similarity of them to aspect will make the control of sentence-aspect similarity model possible. By setting different relevance scores, the accuracy of the aspect-relevant sentences will vary, so that can help find the most-fit one.

**2.2.3 Sentence-aspect result validation**

In this research, the author will select one Airbnb home with a considerable number of reviews. Then, implement the sentence-aspect relevance model to all the reviews, change the relevance score and obtain the list of aspect-relevant sentences. Finally, read through the results and judge the correctness of relevance.

After the model is validated, it can be used for the next step study of this research, the aspect-sentiment analysis.

**Part 3: Sentiment analysis of reviews based on aspect score**

**Part 3.1: Additional step of data selection in the sentiment analysis phase**

Group the selected Airbnb homes by aspect scores and then sample reviews to balance the amount of research data and time limitation of master paper. The segmentation of scores group is acceptable if necessary.

Example: the result of grouping homes by the scores of “cleanliness” is as in the left table:

The total number of reviews of homes with smaller than 7 points of cleanliness are far less than those with no less than 8 points. In this condition, processing all reviews of homes with 9 or 10 points is not applicable and necessary considering the time limitation. Also, according to previous studies of Airbnb reviews, it is validated that Airbnb review score has a positive bias. So, it is a reasonable way to balance the data size for aspect scores that grouping together reviews of homes that have no more than 7 points and meanwhile, randomly select a number of homes from those with no less than 7 points. In the end, the number of reviews of less-than-7-points homes should be as much as possible while the number of reviews of homes with other scores should be kept in around 1500.

**Part 3.2: Determine the sentiment polarity of each sentence**

After regulating the homes-selection of different aspect scores, the next step should deal with the sentiment polarity of sentences. I will use the tool, Stanford CoreNLP, which has the underlying technology based on a new type of Recursive Neural Network that builds on top of grammatical structures.

Compared with other widely-used sentence sentiment analysis tool, such as the NLTK Text-Processing API which uses the lexicon approach, Stanford CoreNLP’s algorithm is more advanced and comprehensive because its learning model actually builds up a representation of whole sentences based on the sentence structure. It computes the sentiment based on how words compose the meaning of longer phrases while the lexicon approach is simply giving positive or negative points for words and compute the sum of all the points which will ignore the order of words and lost some important information.

The Stanford model will provide the sentiment polarity result of sentences in five extents: very positive, positive, neutral, negative and very negative with the sentiment value 4, 3, 2, 1, 0 respectively.

**Part 3.3: Observe the sentiment distribution of reviews for homes in each score groups**

For homes in one score, such as all homes with cleanliness score 8, plot the distribution histograms of relevant-sentences’ sentiment for each home. Then, observe all the distributions to see if there exist a distribution pattern of each score. Or, maybe unexpectedly, there may exist some irrational distribution of positive, negative, etc. which may indicate some problem of Airbnb’s home scoring results.

## Results and Illustrations

**Part 1: Data selection using Normal Distribution**

Airbnb homes defer to Normal distribution in price attribute. Using conf\_intveral and cumulative density function, we can find that the original distribution is a normal distribution because the prob is equal to the confidence\_interval we set as 0.8.

Then, I got the homes selection range: $15 to $243. Actually can be $25 to $243 as the smallest price of all homes is $25.

There are 2890 Airbnb homes and 81275 reviews in total in the selected dataset.

**Part 2: Aspect relevant bag-of-words**

Use the WordNet tool. See sample:

* Aspect: location
* Airbnb description: Location How did guests feel about your neighborhood?
* Keywords from official description: location; neighborhood
* For the word “location”:
  + See all synsets description: synset\_description('location'), result:
    - location.n.01 : a point or extent in space
    - placement.n.03 : the act of putting something in a certain place
    - localization.n.01 : a determination of the place where something is
    - location.n.04 : a workplace away from a studio at which some or all of a movie may be made
  + Only select the relevant synset: “location.n.01”
  + Do the same operation for the word “neighborhood”.
  + Get and store all the synonyms: ['locality', 'neighbourhood', 'neighborhood', 'location', 'neck\_of\_the\_woods', 'vicinity', 'region']

Further description (antonyms usage)

**Part 3: Sentence-Aspect relevance score**

* Explain the principles of Word2Vec and the weighted term frequency method used in the method.
* Explain the relevance score parameter in meaning and use points of view.
* Use a demo to validate this model and explain why this relevance model is ready-to-use for the next part.

**Part 4: The sentiment distribution chats of homes in different aspect scores**

* The model used to determine sentence sentiment: Stanford CoreNLP sentiment model which use the RNN algorithm in the underlying technology.
* Explain the sentence sentiment detection logic and principles.
* Plot the result of each home with the proportions of the sentence by sentiment extents in the histogram chart.
* Observe the patterns of each score if exist.
* Explain some unexpected patterns.

## Discussion

## Conclusion

Conclude from the results. Two parts:

* The sentence-aspect relevance model of Airbnb reviews.
* The sentiment distribution patterns of Airbnb homes according to different aspect scores.

## Future work

## References

* Abdulla, N., Mahyoub, N., Shehab, M., & Al-Ayyoub, M. (2013). Arabic sentiment analysis: Corpus-based and lexicon-based. In Proceedings of The IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT).
* Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.
* Itani, M. M., Zantout, R. N., Hamandi, L., & Elkabani, I. (2012, December). Classifying sentiment in Arabic social networks: Naive search versus naive Bayes. In Advances in Computational Tools for Engineering Applications (ACTEA), 2012 2nd International Conference on (pp. 192-197). IEEE.
* Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15-21.
* Kohail, S. N., & El-Halees, A. M. (2011). Implementation of data mining techniques for meteorological data analysis. Intl. Journal of Information and Communication Technology Research (JICT), 1(3).
* Wei, W., & Gulla, J. A. (2010, July). Sentiment learning on product reviews via sentiment ontology tree. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 404-413). Association for Computational Linguistics.
* Boiy, E., & Moens, M. F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. Information retrieval, 12(5), 526-558.
* Ding, X., Liu, B., & Yu, P. S. (2008, February). A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 international conference on web search and data mining (pp. 231-240). ACM.
* Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.
* Guttentag, D. (2015). Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. Current issues in Tourism, 18(12), 1192-1217.
* Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. International Journal of Hospitality Management, 62, 120-131.
* Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10(pp. 79-86). Association for Computational Linguistics.
* Devaraj, M., Piryani, R., & Singh, V. K. (2016). Lexicon ensemble and lexicon pooling for sentiment polarity detection. IETE Technical Review, 33(3), 332-340.
* Phand, S. A., & Phand, J. A. (2017, October). Twitter sentiment classification using Stanford NLP. In Intelligent Systems and Information Management (ICISIM), 2017 1st International Conference on (pp. 1-5). IEEE.
* Salehan, M., & Kim, D. J. (2016). Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. Decision Support Systems, 81, 30-40.
* Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier. arXiv preprint arXiv:1610.09982.
* Jo, Y., & Oh, A. H. (2011, February). Aspect and sentiment unification model for online review analysis. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 815-824). ACM.
* Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. Expert Systems with Applications, 41(17), 7764-7775.
* Uma Ramya, V., & Thirupathi Rao, K. (2018). Sentiment analysis of movie review using machine learning techniques. International Journal of Engineering and Technology(UAE), 7, 676-681