Tara Blundell

SML310 Final Paper

A Comparative Assessment of Topic Modeling Algorithms using Airbnb Review Data

ABSTRACT

The present study examines the effectiveness of several traditional and modified algorithms in modelling topics of a corpus of short Airbnb reviews. The study found Latent Dirichlet Allocation (LDA) to be the most successful approach, outperforming the modified LDA algorithm designed for short texts. Furthermore, Latent Semantic Analysis (LSA) did not perform well on the review data, despite scoring highly for topic coherence.

INTRODUCTION

The availability of textual data has exponentially increased over the past two decades, emerging in the form of social media posts, including Tweets, status updates, online news and product reviews, as well as many other types of information. One important method of text analytics, especially for extracting thematic information and categorizing text, is topic modelling. A number of topic modelling algorithms are currently used, including Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). These algorithms have been used to model the topics of a variety of text corpuses, including news media text, academic literature and social media data. Much of the data currently becoming of interest to research is web-based and often shorter than the types of texts topic modelling techniques were designed for. Instead of lengthy documents like news or academic articles, Tweets and customer reviews are becoming more available as sources for research (Huang et al, 2014, Sun et al., 2013). Thus, there is an ongoing process of refinement of these techniques for shorter text, especially social media text (Zhao et al, 2019).

The present study seeks to analyze the performance of traditional and modified topic modelling techniques on short text. Although a growing body of research has employed a range of topic modelling techniques to Twitter text (Fang et al, 2016; Zhao et al, 2019), much less research has considered their effectiveness on review data. Another form of short, web-based texts, reviews are valuable to business owners, consultants and marketers as they provide insight into the desires of their customers. Thus, this project seeks to evaluate the strength of various topic modelling techniques in their ability to identify key and cohesive topics of a dataset of online reviews. In this study we will use a large dataset of Airbnb reviews to test both tradition and emerging topic modelling strategies. Furthermore, the study will assess the performance of the various methods for evaluating model quality, as there are many approaches to this, and each has its weaknesses. Three techniques will be considered in this project, as they represent those that have been used most widely or are tailored to short texts. These are, the LDA method, the Augmented-LDA approach, and the LSA method. The topic models will be evaluated using qualitative evaluation, coherence scoring and document classification consistency.

TOPIC MODELLING ALGORITHMS

Latent Dirichlet Allocation

At the moment, the standard strategy used for text topic modelling is the Latent Dirichlet Allocation (LDA), an unsupervised machine learning method that finds latent topics to explain patterns in text data (Reisenbichler & Reutterer 2018). The approach assumes that each document in a corpus contains a collection of topics and that each topic is composed of a mixture of words (Reisenbichler & Reutterer 2018). The number of topics the model assumes exist in the corpus is specified by the user (metrics for choosing appropriate number of topics will be discussed later).The algorithm finds these topics in a corpus of documents by identifying words that are often found in the same document and merging them to generate a topic (Reisenbichler & Reutterer 2018). One way to think about the technique is representative of a generative process for documents in a corpus (Hong & Davison, 2010). For each document there is a multinomial distribution of topics and, within each topic, a distribution of words associated with it. A process of selecting a topic and a word from within the topic is repeated to fill all words in a document. The distributions of topics and words for each document represent the output of the model. The probabilistic distribution of topics for each document provides estimates of how probable it is that document belongs to a topic. Therefore, for one document probabilities for each topic sum to one. These probabilities can be used to categorize new documents into the topic which the model indicates it is most likely to belong to (Hong & Davison, 2010).

As mentioned, an important feature of the LDA technique is that it relies on the pattern of terms within individual documents to find topics. For example, if two terms frequently appear in the same document, they are likely to be included in the same topic, as the model infers that they are closely related. Because LDA is a "bag of words" model, the order of the words in each document to not matter (Reisenbichler & Reutterer 2018). Thus, the typical input for an LDA is a document-term matrix, in which each column represents a word from the corpus and each row represents a document. In each cell, is the value for the number of times the corresponding word appears in the corresponding document (Reisenbichler & Reutterer 2018). Clearly, the ordering of the words is not relevant, but the model relies how terms cooccur within documents to identify which words are related and therefore likely to be part of the same topic. As a consequence, document sizes have an impact the number of word-cooccurrences. In corpuses with smaller documents, words overall cooccur less frequently than when the corpus contains larger documents. Because the LDA algorithm was created to be used with large documents, it often underperforms on text corpuses with short documents (Zhao et al, 2019), like data from social media or product reviews.

Augmented LDA

One approach to enabling effective use of LDA for short texts is by combining them into larger bodies along some important category. This is known as the augmented LDA (aLDA) approach and is used to increase the co-occurrences of relevant words within a single document allowing the LDA method to better detect word patterns on the document level (Zhao et al, 2019). The method has been used in previous research on Twitter data, whereby tweets by the same author are combined into longer documents, which were then modelled using LDA (Zhao et al, 2019). One study which took this approach compared the results of approach with the traditional LDA approach, which used individual tweets as documents. In order to assess the performance of each approach, they compared the coherence scores, a common measure of the quality of learned topics, of topics generated by each technique (Zhao et al, 2019). The coherence score evaluates the semantic similarity of terms within each topic, with better scores attached to topics composed of words with close semantic associations (Fang et al, 2016). They found that for a dataset of tweets, augmenting into larger documents before carrying out the LDA produced topics with higher coherence scores. Another study sorted microblogging posts into larger documents based on the issue they addressed, before topic modelling using LDA (Hajjem & Latiri, 2017). This study also found that the LDA method produced more coherent topics when used on augmented documents rather than the individual text pieces (Hajjem & Latiri, 2017). These studies indicate that that the augmenting text documents based on some meaningful categorization help improve the performance of the LDA method for traditionally short text data.

Latent Sentiment Analysis

Another widely used approach to topic modeling, is based on latent sentiment analysis (LSA), which had a foundational impact on the field of topic models (Cvitanic et al, 2016). Compared with LDA, the LSA method is a more basic approach that was not specifically developed for topic modeling. It is used for many tasks in Natural Language Processing, especially for assessing similarity between documents or words. It is often used for data clustering, document classification and finding synonyms of terms, but can also be applied to topic modelling. Similar to LDA, the process begins with a document-term matrix and considers the association of words on a document level to identify topics (Bergamaschi et al., 2014). However, unlike LDA, this does not use a generative method, so it does not create probabilistic models of documents and topics. Instead, it takes an entirely different approach, using single variable decomposition (SVD) to factorize the document-term matrix and reduce its size (Bergamaschi et al., 2014). Through this process it creates a high-dimensional semantic space that can be used to compute similarity between terms, sentences and documents (Cvitanic et al, 2016). The measurements of semantic similarity of terms in this space and their patterns within documents are then used to construct topics. Thus, instead of modelling a probabilistic generative process like LDA, LSA uses assessments of semantic similarity to group terms into topics (Cvitanic et al, 2016). Using this simpler approach, it finds topics more quickly than the LDA algorithm. However, because does not use model probabilistic distributions of terms and topics, many researchers believe that the topics it produces do not model the data as effectively as the LDA method (Cvitanic et al, 2016). Empirical studies disagree on whether this is the case.

Researchers have compared the LSA method with the more commonly used LDA method for topic modelling. One study aimed at generating topics from railway accident text data compared the performance of these two approaches (Williams & Betak, 2018). The goal was to classify accident reports into topics regarding the nature of the accident. Both of the models identified many of the same topics, while each one produced topics that the other didn't. The researchers indicated that the techniques therefore performed similarly well, however, their evaluation of the topics was limited to this basic qualitative analysis (Williams & Betak, 2018). Another study used the two techniques to identify topics in a database of movie plots, in order to classify movies and provide movie recommendations (Bergamaschi et al., 2014). The researchers evaluated the models by comparing how they classified movies into topics with how movies were classified by human participant and on websites like IMDb. They found that the LSA method performed better in classifying similar movies into the same topics (Bergamaschi et al., 2014). Perhaps this result is not surprising as the LSA model is more widely used for classification of documents than the LDA, which was designed to produce more coherent topics. This issue brings rise to the question of how topic models should be evaluated, which is a key consideration of this paper. There are numerous metrics on which to evaluate topic models, but the most appropriate or informative analysis will depend on the goals of the project. As the studies described in the above examples sought primarily to classify documents into topic clusters, accuracy of classification may have been the most important metric for evaluation.

TOPIC MODEL EVALUATION

Qualitative Evaluation

In many of the studies analyzing the performance of topic models and the validity of evaluation measures, the human evaluation of topics is the standard to on which to base conclusions or compare other measures. However, it is costly to ask a group of participants to rate topic quality for various models, especially when each model involves a large number of topics. In previous studies, researchers have recruited participants through crowdsourcing platforms to be trained to recognize quality topic and then rate the topics of models created by different approaches (Fang et al, 2016). In one study, the participants consider the top 10 words in two topics and compare how accurately they summarize the topic of a given tweet (Fang et al, 2016). Given the cost of hiring and training participants this method is often not viable for many researchers. The present research will use an adapted version of human evaluation by which the researcher themselves evaluated topic models based on a predefined set of qualitative criteria. This method is versatile as the criteria can be altered depending on the goals of the project. In this project, we rely on previous research to guide the criteria of importance for topic quality (Fang et al, 2016). These criteria are:

1. Overlap between topics is minimized

    a. There is a lack of repeated words between different topics within a model

2. Topics are informative regarding types of issues customers include in their reviews

    a. There is a lack of general words in the topics

b. Each topic focuses on specific and closely related themes

The defining of specific criteria minimizes the subjectivity of qualitative evaluation and tailors the measurement of topic quality to the specific goals of the individual model. Although human evaluation of topics based on qualitative criteria can be costly or subjectivity, it remains an important standard to which to compare the performance of other evaluation methods.

Topic Coherence

Topic Coherence scores are the most prominent measure for the quality of topics produced by topic models. Measuring coherence of topics involves quantifying how closely associated the words are in each topic. The coherence can be measured using several different measuring approaches, but the most commonly measure is based on how frequently they cooccur within documents (Lau et al, 2014). This is an important factor for topics because it shows how well they reflect the patterns of word use within the corpus. For example, if a topic is comprised of words that rarely occur together in the text corpus, this topic is unlikely to provide any information about the structure of the data. The range of the scores depends on the specific method in which it is measured, which can vary somewhat as there are several ways to implement the analysis (Lau et al, 2014; Stevens et al, 2012). One common method is using the UMass metric, which as mentioned above, computes cooccurrences of words within the same document of the training corpus (Stevens et al, 2012). Equation 1 shows the metric is computed, with *D(x, y)* equal to the number of documents containing words *x* and *y* and *D(x)* the number of documents containing *x* (Stevens et al, 2012). The closer the UMass value is to zero, the more coherent the topic it measures (Mimno et al, 2011).

$$score(v_i, v_j, \epsilon) = \log \frac{D(v_i, v_j) + \epsilon}{D(v_j)}$$

Equation 1. Formula for UMass Coherence (Mimno et al, 2011)

A recent study tested many different coherence metrics and found the C_V coherence metric to be the most consistent with human appraisal of topic quality (Syed & Spruit, 2017). It combines several measures, including one similar to the UMass metrics, which calculates how frequently similar words appear in the same document. It also considers word pairs, whereby each word is paired with every other word in the text. The word and word pair probabilities are compared to determine the relatedness of the two words. This measure is used to assess how associated word are within each topic and thus calculate the overall coherence of each topic (Syed & Spruit, 2017).

In evaluating a model with more than one topic, the coherence scores of each topic are calculated and then averaged to measure the performance of the model as a whole. A model’s mean coherence score generally improves as the number of topics in the model increases. As previously mentioned, the user can decide how many topics they would like the model to find in the corpus. In models with fewer topics, the topics tend to be broader and potentially encompass numerous sub-topics. With more topics, they are generally more specific and therefore the words within the topic tend to be more closely associated, leading to better scores for coherence. A study that used coherence scores to measure the performance of an LDA model on scientific publications found that coherence score increased greatly as the number of topics increased from 1 to 20, before plateauing for models with more than 20 topics (Syed & Spruit, 2017).

Although a model with more topics may have a higher average coherence score, depending on the purpose of the model, this may or may not be desirable. If a researcher is interested in extracting more broad topics from their data, a lower coherence score may be preferred. Coherence scores do not therefore represent a universally useful method for assessing topic quality but must be considered on the basis of the goals of the analysis. Nevertheless, the coherence score metric can be useful for comparing across models each with the same number of topics. In this project, it will be used to compare the quality of topics between models of the same size, that were produced by different modelling techniques; thereby comparing the performance of the techniques in modeling the given data. A previous study has already used coherence scores, specifically the UMass measure, to assess the quality of topics in models produced by the LDA and LSA methods (Stevens et al, 2012). It found that average topic coherence for the LSA model was higher than for the LDA model when there were few topics in the model. However, as the number of topics increased, the average coherence for the LDA model improved above the LSA coherence, which declined as the number of topics increased (Stevens et al, 2012). The coherence metric is one of those used in this study to evaluate the models. It will be interesting to note whether these trends appear in the results of the present study.

Document Classification

Another metric for evaluating the topic models is by testing how accurately they are able to classify data. This method requires that the researcher identify features of the data upon which it may be most meaningful to classify the data. For example, if a researcher may know that each of the documents in a corpus belongs in one of five possible categories. They would create a topic model with five topics and would hope that each topic reflects a distinct category and that all of the documents in that category be classified under the same topic. This outcome would reflect that the model found topics that represented the already known categories, and thus performed well. This simple example demonstrates how classification accuracy can be used to assess topic models. This technique has already been used in the literature. One study was designed to use this method for evaluating topic models of Tweets, as Tweets were extracted from the database to clearly fall under one particular topic (Fang et al, 2016). They found a modified LDA algorithm to be more accurate for classifying the tweets than the traditional LDA algorithm (Fang et al, 2016). In practice however, there may not exist clear, distinct categories into which documents fall. Furthermore, even if the corpus does divide naturally into categories, it is not always clear that a given topic models one particular category, making it difficult to evaluate accuracy with the traditional document classification approach.

A novel approach to assessing document classification accuracy will tested in the present the study. This novel approach addresses the uncertainty around whether each topic generated by the model actually reflects each known category of the corpus. Instead of assuming one category is associated with a give topic, this measure assumes that documents in the same category should have a similar likelihood of being placed in a given topic. Thus, instead of binary topic assignment, this method considers the likelihood of that a document would fall under given topic. The assignment probability takes into account both the words in the test document and the most prominent words in the topic to calculate the likelihood that the document's content fits with the topic. As mentioned, a good model will assign documents of the same category to a given topic with a similar probability. In order to calculate the probability similarity across many documents of the same category, the variance of the probabilities is calculated. Using this method, the overall classification accuracy of a model can be calculated by averaging these variances across all document categories and all topics in the model. These mean variance measures can be used to compare models trained and tested on the same data, in terms of how consistently they assign a document of the same category to the same topic.

DATA

General Information

The Airbnb review data was freely available on the Inside Airbnb database. There were separate datasets for a number of cities across the US and other countries. Only data from the United States was used. The datasets from seven cites were included in the project, each containing between 2,500 and 50,000 listings. The cities were: Austin, TX; Boston, MA; Chicago, IL; Denver, CO; Fort Lauderdale, FL; Las Vegas, NV; Seattle, WA. There was a listing dataset for each city included details about the properties including overall rating, location and size and nightly rental price. Thus, in the raw dataset, there were 62,494 listings. There was some data missing, most notably in the columns for zip code, square feet size and listing ratings. Only data with reviews missing were removed. Reviews were missing for 12,000 listings, many of these properties were new to the site or for another reason had not yet been reviewed. These listings were not valuable for the project and were removed, after which 50,400 listings remained.

The reviews constituted a second dataset, which could be associated with the listing dataset through a common listing id variable. A numerical rating associated with each review was not provided, only the overall rating for the listing averaged across reviews. Given that there were at least 5,000 reviews available for each city, the final review dataset contained a random sampling of 5,000 reviews from each of the seven cities. In many instances, multiple review instances in the final review dataset were related to the same property, but this was not considered a problem for the purpose for this research. Text data was tokenized and stemmed before data exploration.

Exploratory Data Analysis

The Exploratory Analysis was conducted in R. Several additional variables were explored along with the review data in order to gain more information about data. Furthermore, the exploratory analysis sought to identify variables on which the reviews could be categorized. This was considered important, firstly, as it was necessary to identify along which variable reviews should be combined into larger documents for the aLDA method. Secondly, it is necessary to find a meaningful way to categorize the data that would allow for the models to be evaluated in terms of their accuracy of document classification. This was identified as one of the key metrics for model evaluation.

Price was considered to be one of the most important and informative variables in the dataset. The price is measured as the nightly cost in USD charged by the owner to lease the accommodation. The prices ranged from 0 to $24,999 per night across the dataset. Large outliers were excluded, and Figure 1 shows the price distribution of the prices of listings between 0 and $1,000 per night. The distribution has a positive skew. Given that the distribution was skewed the median was used as the measure of central tendency. The median daily price of across listings is $115. The median price for each state was plotted in a bar graph in Figure 3. This showed that states have similar median prices.

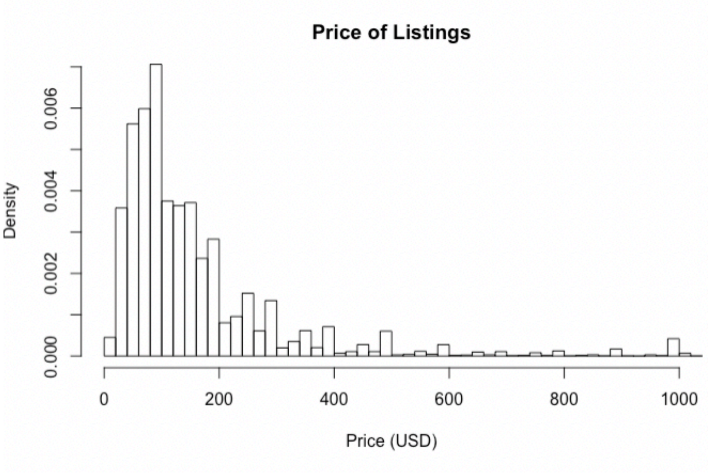

Figure 1. Histogram of listing prices below $1000

Property type included the following levels: apartment, house, guest suite, cabin and bed and breakfast. A plot of the main property types (those with greater than 200 listings) in the dataset is shown in Figure 3. The plot shows that house, apartment, condominium and guest suite are the most common property types in the dataset. The other categories were much less common, indicating that this variable may not be a useful on for measuring classification accuracy.

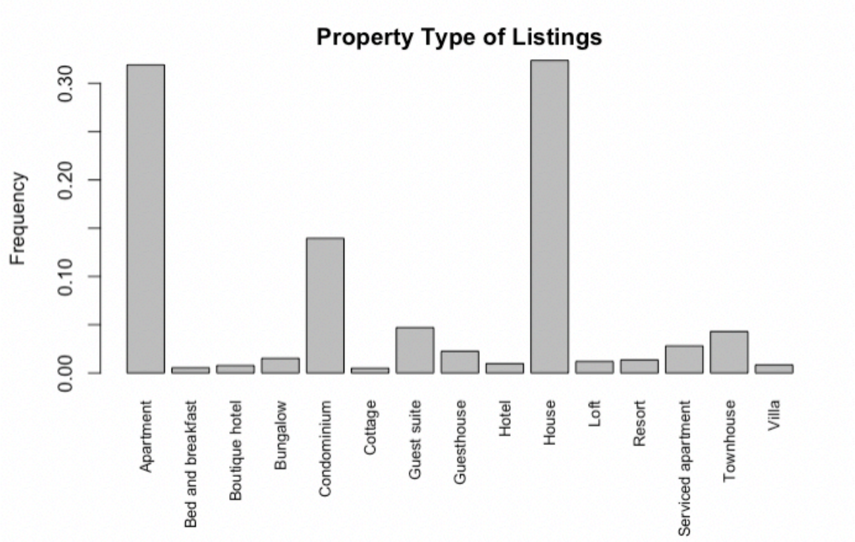

Figure 3. Proportion of listings for main property types

One other important variable in the dataset is average overall rating for each listing. Ratings are measured on a scale of 0 to 100 and in the dataset ranged from 20 to 100. Ratings were also given for more specific issues, including location, cleanliness, communication and value. The histogram of overall ratings shown in Figure 4, reveals a negative skew in this distribution, in which the majority of ratings were close or equal to 100. Clearly, the most frequent rating was 100, and the median was 97. The median rating across cities were not significantly different, ranging between 96 and 99. This indicates that Airbnb reviewers tend to give high ratings for their accommodation. This is relevant to consider when analyzing the reviews, as it indicates that reviews are likely to tend towards being positive and flattering.

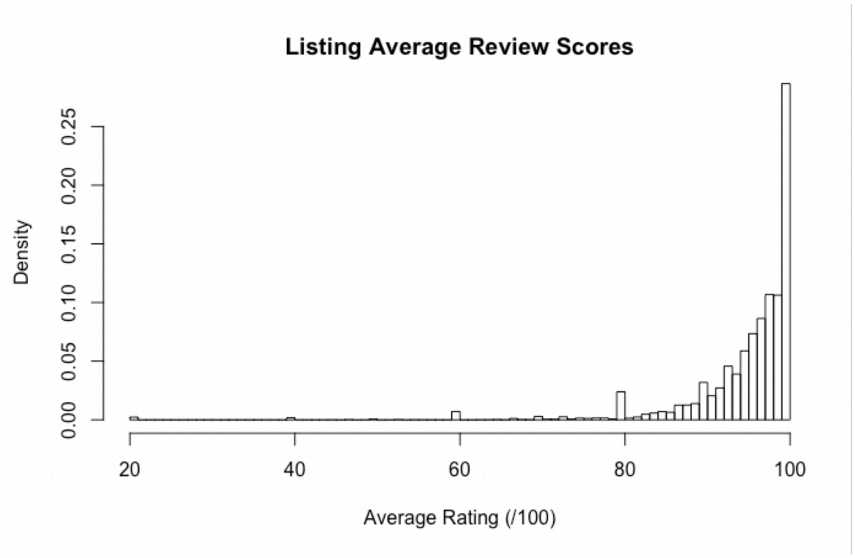

Figure 4. Histogram of listings' average review ratings out of 100

Given that the explored variables in this dataset were generally skewed, including price, review score, and property type, these were not considered to be viable options to split the data into categories. Given that the review dataset was already evenly divided into seven cities, it seemed most obvious to use this as the variable on which to divide the data into categories. Thus, location was used to divide reviews into larger documents for the aLDA model. These categories were also used to assess the consistency of document classification as a third metric for model performance.

The data of primary interest in the dataset are the review comments. Figure 5 shows the frequencies of the 20 most prominent terms in the reviews. These include many general terms related to an accommodation stay, which are not particularly informative about the topics of importance to guests; these include, “stay”, “house”, and “recommend”. However, other frequent terms may indicate some of the aspects of the experience that are important to customers, including “location”, “host”, “clean” and “quiet”.

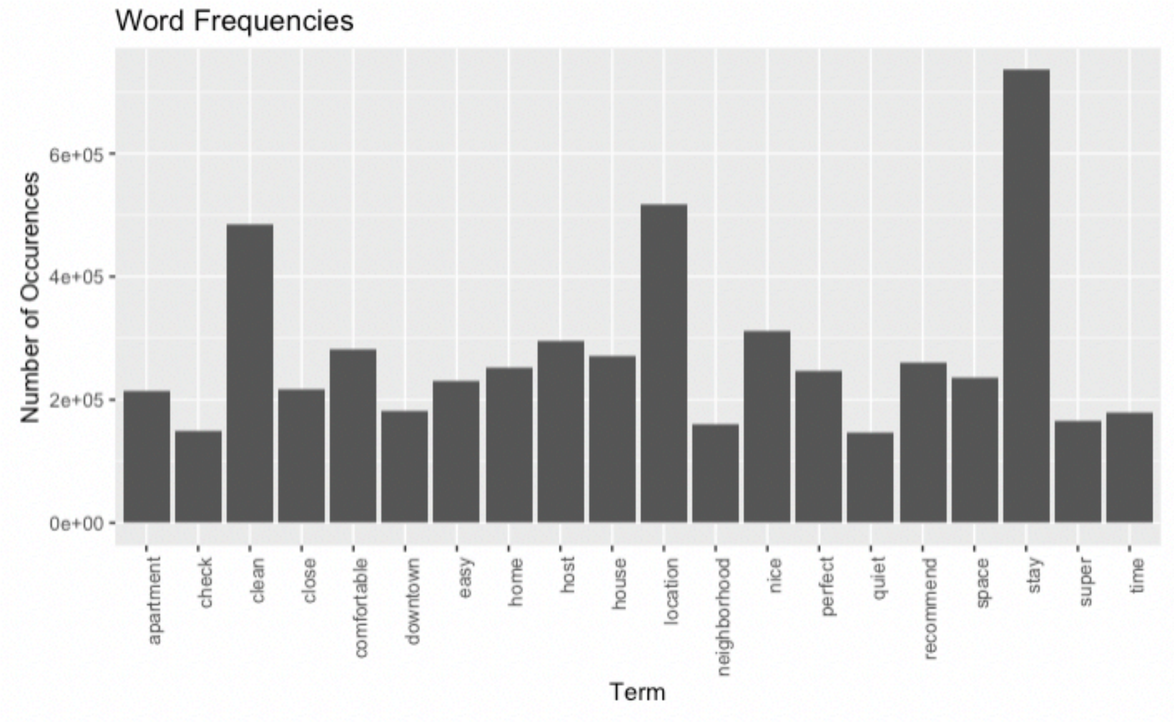

Figure 5. Frequent terms in reviews

The frequently used bigrams (two words which appear together in the text) are shown in Figure 6. This revealed further information about guests' priorities, including "walking distance", "super clean", "easy access" and "quiet neighborhood". These analyses indicate some of the topics that may be found through the topic modelling, including the property's convenience, noise and cleanliness, and the helpfulness of the host.

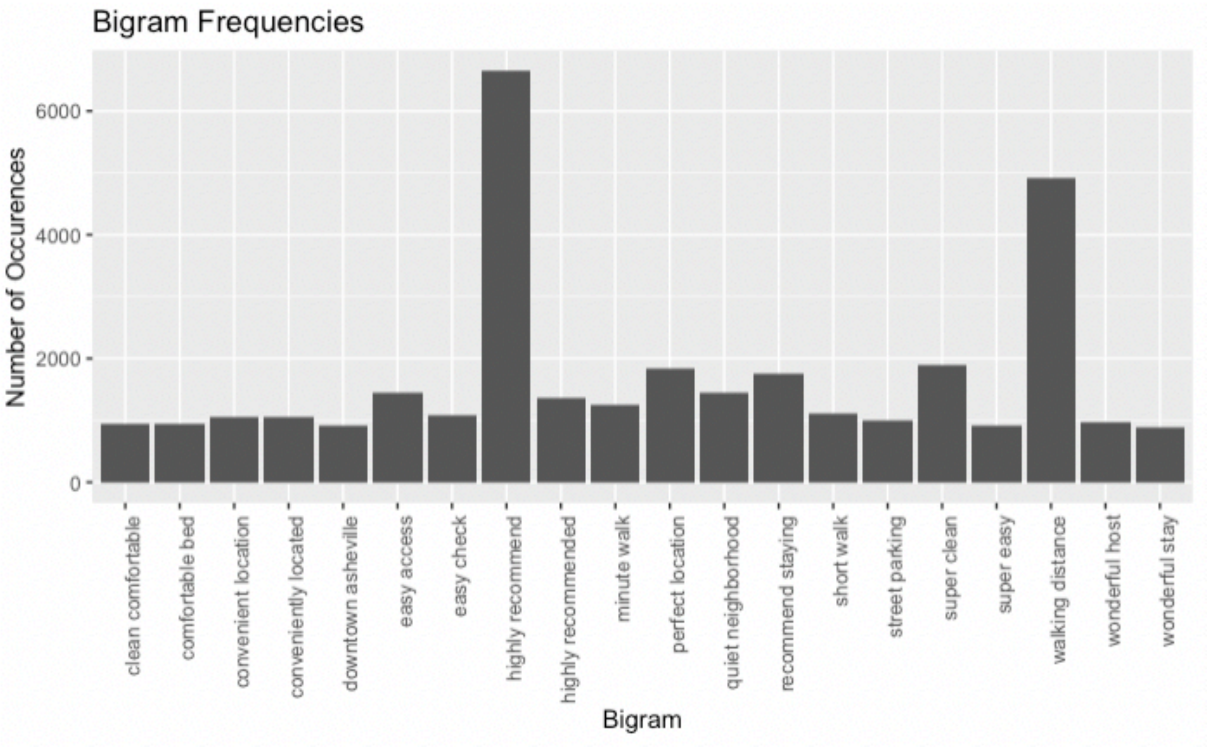

Figure 6. Frequent bigrams

METHODS

The data preparation and analysis were conducted in R. After the review data was loaded into R, each review was tokenized, any stop word or tokens containing numbers were removed and remaining words were stemmed in preparation for analysis. Tokens that occurred in fewer than 10 reviews or that constituted more than 20% of all words were excluded from the analysis. After this processing, the dictionary of tokens across all reviews contained 5,616 distinct tokens.

After this text processing, the text corpus remained in distinct documents, based in individual reviews. The reviews were split into training and test sets, which comprised 70% and 30% of the original text respectively. For the tradition LDA and LSA algorithms, the data was ready for model training at this stage. For the aLDA model, the reviews for all properties in each of the eight locations were combined into eight large documents before the models were built.

The three topic modelling methods discussed in Section 2 were implemented in R on the training corpus of reviews. The 'textmineR' package is used to implement the LDA and LSA methods, which automatically generates coherence scores. The models will be trained, each with seven topics, as there are seven cities in the analysis and therefore seven supposed categories. The quality of the models first be assessed using qualitative analysis and coherence metrics. Additionally, document classification consistency will be calculated based on variance in probabilistic document assignment. Document assignment is only available for the LDA model in R. Therefore, only the LDA and aLDA will be tested using this metric.

RESULTS

Qualitative Analysis

A final model was trained using each algorithm, each produced 7 topics. The LDA and aLDA were trained over 400 iterations, as this was optimal for coherence. The first stage of analysis involves comparing the topics produced by these models qualitatively. Table 1 shows the five most prominent terms or bigrams for each topic of each model.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Traditional LDA Model						
room	great	great	stay	apartment	home	house
place	downtown	place	place	close	stay	great
kitchen	walking	host	clean	great	beautiful	group
bed	distance	highly	stay	neighborhood	house	perfect
nice	restaurants	recommend	coffee	walk	wonderful	host
Augmented LDA Model						
austin	place	nashville	apartment	close_subway	las_vegas	stay
strip	ny	denver	room	apartment	casino	host
great	apartment	great	house	room	manhattan	subway
place	chicago	downtown	boston	host	palm	beach
stay	stay	house	seattle	blue_line	mgm	clean
LSA Model						
great place	great_place	good	excellent	excellent	great_location	great_place
great	great	good_place	arrival_automated	excellent_place	location	awesome
place	place_stay	good_location	automated_posting	excellent_location	great	place
good	place	good_stay	automated	excellent_host	spot	excellent
nice	great_stay	place_good	cancelled	excellent_stay	great_stay	place_stay

Table 1. Topics Top Terms

Based on the qualitative criteria outlined in the introduction, this analysis will consider how distinct the topics are and how specific. In each topic model, many topics contain the same common words found in the reviews, including “good”, “stay”, and “nice”, which reduces the distinctiveness of each topic. The LSA topics appear to have the most general terms and thus has less informative topics. The aLDA topics appear to use the fewest general terms and yet does not appear to have found distinct topics. In terms of specificity, the models also differ. The topics of the LSA appear to be somewhat specific, but in a way that focuses on word groups rather than themes. The aLDA has more specific terms but does not appear to combine them in a way that makes sense thematically. The LDA topics appear more informative, as they can to some extent be distinguished from one another somewhat in terms of theme. A more specific difference of not is that the aLDA results appear to include more terms specific to locations.

Coherence Score

The next method of evaluation considered is coherence score. The augmented LDA does not produce valid coherence scores, as there are too few documents over which to compare the term frequencies. Coherence scores for the remaining topic models were assessed with various number of topics. This range was chosen to be 10 to 80 for this analysis, based on previous research that uses a similar range for this analysis (Syed & Spruit, 2017; Stevens et al, 2012). Furthermore, this range was considered broad enough as more than 80 topics would not be useful for this project as information on so many topics would be difficult to interpret and use.

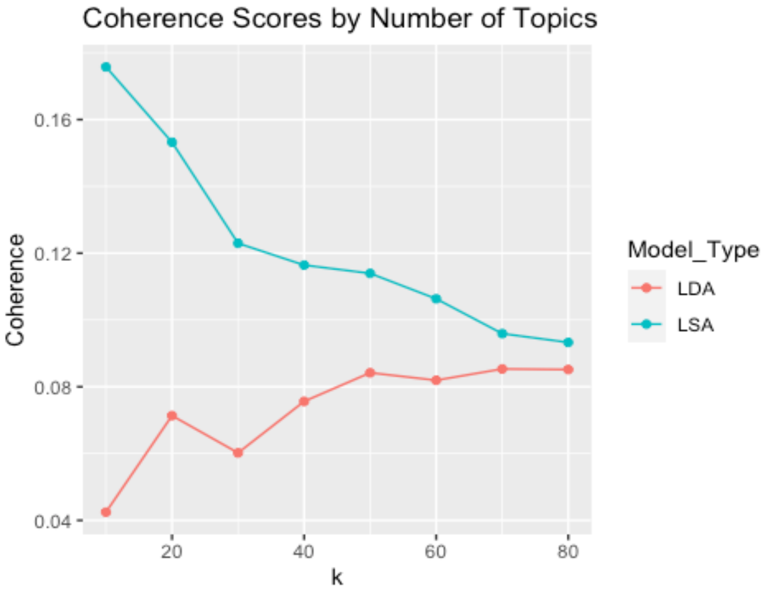

Figure 7. Coherence Scores for Topic Models

The results of this analysis are shown in Figure 7 and support findings of previous research (Stevens et al, 2012), which suggests that the LSA model performs better than the LDA model at lower levels in terms of coherence. However, this result also suggests that the coherence score may not be a reliable metric for indicating how the quality of topics. For example, the coherence score for the LSA method with 10 topics is very high. However, the breakdown of this model was shown on the previous page and was found to be very uninformative and did not appear to map the breadth of themes in the corpus. Thus, we turn to another metric for evaluating the LDA models, which appear to be the most informative based on the qualitative analysis.

Document Classification

An adaptation of the classification method described in the literature review is used here. Given that these topic models "classify" documents into topics with a probabilistic distribution rather than a direct assignment, this approach will test how much this probabilistic distribution varies among documents of the same type. The Exploratory Data Analysis suggested that we one important category in the dataset is state location. We use this in order to understand the actual category each review. If the techniques were modelling topics that accurately represented characteristics of the reviews and listing, for instance the location, the reviews from the same location would ordinarily be sorted into the same topic. Thus, this method calculates the variance between assignments of documents of the same location to the topics. A low variance would indicate that the model represented the location characteristics well, possibly indicating a good quality topic.

The results above show that the augmented LDA method has significantly lower variance than the traditional LDA method in classifying new documents (p-value: 5.277e-13). This indicates that the augmented LDA method is more likely to classify reviews from the same location into the same topic. Thus, the augmented LDA appears to perform better in modeling the aspects of the locations in the topics. This is important, as it shows that this augmented LDA method should be used over the traditional LDA method if the goal of modelling is to reveal information about important themes in different categories. This can of course be applied to other categorical variables, including (for this dataset) price and property type, by augmenting the documents based on the desired category.

DISCUSSION

These results raise several interesting insights and opportunities for future directions. Firstly, is seems clear from this analysis that various techniques should always be used to evaluate topic models. The analysis showed that the widely relied-upon techniques like coherence evaluation do not necessarily indicate that models capture meaningful topics. High coherence value may indicate very specific topics rather than those well representative of the corpus of texts. The LSA model proved to be an example of this, as the topics were highly coherent but failed to be informative regarding the scope of the reviews. More research should be conducted to follow up on this finding, including comparing ratings of topics given by a group of trained participants to several measures of coherence for LSA and LDA models.

Furthermore, a follow-up study should be conducted to further evaluate the relative performance of LDA and aLDA on various types of short text. Using the same dataset, the study should be run again using different categories for augmenting texts. Perhaps zip code could be used to provide a larger number of documents to the model whilst still using location as the variable of categorization. Furthermore, different sources for corpuses of short text should be tested as well, especially those which fall naturally into categories. For example, Tweet data could categorize based on the keyword used to retrieve the Tweets. This categorization may better reflect the content of the texts, compared with location in Airbnb reviews, and therefore enable to the aLDA algorithm to perform better. This recommendation is made with the knowledge that Tweet data is becoming an important source of political, social and economic information. As the field of text analytics grows and the use of short text expands, the investigation of methods for topic modelling and evaluation become increasingly important to enable researchers to gain insight from data and classify new text.

REFERENCES

Bergamaschi, S., Po, L., & Sorrentino, S. (2014). Comparing Topic Models for a Movie Recommendation System. *Proceedings of the 10th International Conference on Web Information Systems and Technologies*, 172-183.

Fang, A., Macdonald C., Ounis, J. & Habel, P. (2016). Topics in Tweets: A User Study of Topic Coherence Metrics for Twitter Data. *European Conference on Information Retrieval*, 492-504.

Hajjem, J. & Latiri, C. (2017). Combining IR and LDA Topic Modeling for Filtering Microblogs. *Procedia Computer Science*, 112, 761-770.

Hong, J. & Davison, B. (2010). Empirical Study of Topics Modelling Technique. *Proceedings of the First Workshop on Social Media Analytics,* 80-88.

Huang, J., Rogers, S., & Joo, E. (2014). Improving Restaurants by Extracting Subtopics from Yelp Reviews. *Social Media Expo 2014*.

Jiang, J. & Conrath, D. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of International Conference Research on Computational Linguistics*, *19-33.*

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence of topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272.

Reisenbichler, M. & Reutterer. T. (2018). Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, 89, 327-356.

Feng-Tso Sun, Martin Griss, Ole Mengshoel & Yi-Ting Yeh. 2013. Latent Topic Analysis for Predicting Group Purchasing Behavior on the Social Web. *Proceedings of the 2013 Uncertainty in Artificial Intelligence Application Workshops*, 67-76.

Lau, J. H., Newman, D. & Baldwin, T. (2014). Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539.

Williams, T., & Betak, J. 2018. A Comparison of LSA and LDA for the Analysis of Railroad Accident Text. *Procedia Computer Science*, 130, 98-102.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. 2012. Exploring Topic Coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952–961.

Syed, S. & Spruit, M. 2017. Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*.

Zhao, Y., Qiao, Y., He, K. (2019). A novel tagging augmented LDA model for clustering. *International Journal of Web Service Research*, 16(3), 19.