Department of Computer Science

MSc Data Science and Analytics

Academic Year 2017-2018

AN ASPECT-BASED SENTIMENT ANALYSIS OF CUSTOMER REVIEWS

Issam Jaber 1740856

A report submitted in partial fulfilment of the requirement for the degree of Master of Science

Brunel University
Department of Computer Science
Uxbridge, Middlesex UB8 3PH
United Kingdom
Tel: +44 (0) 1895 203397

Fax: +44 (0) 1895 251686

Abstract

With the increase in customer reviews being published, mainly due to the rise of e-commerce, there is a demand for adequate methods to extract meaningful information from the data produced. Thus, in this dissertation, we propose a novel approach to Aspect-Based Sentiment Analysis (ASBA) that is capable of accurately retrieving customer sentiments regarding specific aspects through analysing customer reviews. An ABSA is a technique that attempts to discover the most important aspects within a textual document and classifies the sentiment polarity of the discovered aspects. Numerous ABSA methods have been proposed in the past, however, the majority of the models proposed were not scalable, and were mainly domain specific. The approach needs to be robust and versatile with the ability to perform across domains and languages, this is because customer reviews are being produced for a variety of products and services. ASBA is generally composed of two tasks, Aspect Detection, and Sentiment Analysis [Schouten and Frasincar, 2016]. Aspect Detection is the process of detecting aspects of an entity in a textual document. There are two main approaches to Aspect Detection – supervised and unsupervised. Sentiment Analysis captures the opinions and attitudes conveyed in text [Liu and Zhang, 2012]. For the Aspect Detection phase, we presented a Noun Only Approach to the topic modelling technique Latent Dirichlet Allocation (NOA-LDA). We found that due to the unsupervised nature of the approach, the approach was versatile and capable of performing accurately across domains. We also found that the NOA-LDA system was superior to a raw corpus LDA, both in terms of coherence and scalability. For the Sentiment Analysis phase, we presented a Pragmatic Lexicon Scoring System (PLSS). Our method uses a corpus of customer reviews to produce an opinion Lexicon list.

Acknowledgements

I wish to express thanks to my supervisor Dr Yongmin Li, who's guidance and expertise assisted me profoundly throughout this dissertation. The frequent meetings we had, and quick replies provided me with the support I needed to complete this dissertation.

I would also like to express thanks to my family, who have been patient and understanding whilst I was completing this dissertation.

I certify that the work presented in the dissertation is my own unless referenced

Signature: Issam Jaber

Date: 21/09/2018

TOTAL NUMBER OF WORDS:

10,130

Tables and Figures

TABLE OF CONTENTS

Chapter 1: Introduction

Modern consumers are increasingly dependent on online resources to assist them in making purchasing decisions. The main resource being online customer reviews. A customer review is a critical assessment of a product or service purchased, used, or experienced by the customer. It has rapidly become the norm for e-commerce sites, such as Amazon.com and eBay.com, to provide customers with the option of leaving feedback regarding a product or service they purchased.

A recent survey by Podium [2017] states that 93% of consumers say online reviews do influence their purchasing decisions. With such a vast array of consumers influenced by online reviews, reviews are rapidly becoming the most significant resource for generating new customers. Studies have even gone on to state that online customer reviews are having a larger impact on profits than traditional media [Pang and Lee, 2008]. Customers are more drawn to online reviews due to the convenience and credibility. Online reviews are credible because customers have no reason to lie when leaving them, this is because the online aspect makes them anonymous and impersonal [Stephens-Davidowitz, 2017]. Whereas, paid marketing is owned by the business and has every reason to exaggerate. Recent surveys have also suggested that 85% of consumers trust customer reviews as much as personal recommendations [BrightLocal, 2017].

There are certain properties of online reviews that influence a customer's decision. The main being the overall rating, and the number of reviews. A high overall rating provides customers with an immediate quality approval, a large number of reviews certifies the overall rating. The latter is supported by a recent study, the study found that given two products or services with similar overall ratings, consumers are more inclined to purchase the product or service with the larger quantity of reviews [Powell et al., 2017]. As a product or service appreciates, online reviews become increasingly significant. This is due to the increased risk involved with making a bad investment.

Considering that online reviews are beneficial towards customers, they could also be highly beneficial towards businesses. A study found that a one-star increase to an overall rating will lead to a 5-9% increase in business revenue, whilst a single bad review could cost a business 30 customers [Luca, 2011]. Furthermore, the study found that customers are likely to spend 30% more if a business has excellent reviews. The question is, ‘how does a business improve its rating?’.

In order for a business to improve its overall rating and reviews, they must know what aspects of the product or service customers feel need improving. Most review systems utilise the comment and rating system. The rating only provides a business with an indication of the overall customer sentiment towards the product or service, it does not specify what contributes to the rating. In essence, the rating is capable of telling a business that their product or service needs improving, but it does not identify what precisely about the product or service the business needs to improve. The answer usually lies in the comments; a recent survey found that ‘73% of consumers value the written review over the overall star rating’ [Fan and Fuel, 2017]. Thus, a business should also value the comments. The comments are usually more detailed, and they convey more than just sentiment. Comments reveal what customers think, and why they think it. Every product or service generally consists of several aspects. For instance, a Camera will have aspects such as zoom, megapixels, screen size, etc.

A customer review will usually address at least one aspect. Knowing what customers feel regarding certain aspects can benefit a business in a way that ratings cannot. Businesses can utilise this information to improve certain aspects of their product or service. Likewise, a business can discover the aspects that are most popular with their customers, enabling them to market the product more effectively. Thus, analysing the customers' comments could provide a business with vital information regarding their product, giving them a competitive advantage. There is a problem, however, that is ‘how should the comments be analysed?’.

A business could simply read their reviews, and make a conclusion that way. The issue with this is: ‘what if a business has thousands of reviews for a product or service?’, or ‘what if the business offers many products or services?’. This could make it very difficult and time-consuming for a business to analyse the reviews. For instance, a company like Sony likely has hundreds of products on sale with each product having a number of reviews ranging from the hundreds to the tens of thousands; making it incredibly inefficient to manually analyse each products' review comments. Even if a business does not have a large number of customer reviews, figures show that annual e-commerce sales have grown by an average of 25% year-on-year since 2014 [Statista, 2018]. As the number of consumers shopping online increases, the number of customer reviews being published should also increase. The need for a computational technique capable of retrieving customer opinions regarding specific aspects through analysing customer reviews is vital in today’s competitive landscape.

In this dissertation, we propose a novel approach to Aspect-Based Sentiment Analysis (ASBA) that is capable of accurately retrieving customer sentiments regarding specific aspects through analysing customer reviews. An ABSA is a technique that attempts to discover the most important aspects of a textual document and classify the sentiment polarity of the discovered aspects. Numerous ABSA methods have been proposed in the past, however, the majority of the models proposed were not scalable, and were mainly domain specific. The approach needs to be robust and versatile with the ability to perform across domains and languages, this is because customer reviews are being produced for a variety of products and services. ASBA is generally composed of two tasks, Aspect Detection, and Sentiment Analysis [Schouten and Frasincar, 2016]. Aspect Detection is the process of detecting aspects of an entity in a textual document. There are two main approaches to Aspect Detection – supervised and unsupervised. Sentiment Analysis captures the opinions and attitudes conveyed in text [Liu and Zhang, 2012].

1.1 Aim and Objectives

The main aim of this dissertation is to produce a novel ABSA approach for customer reviews that is scalable, versatile, and that produces coherent results. In order to achieve this aim, several objectives must be met, such as:

1. Evaluate the current literature surrounding ABSA to identify the flaws in current approaches.
2. Develop an approach for Aspect Detection that can identify coherent key aspects from a given opinionated text.
3. Develop an approach for Sentiment Analysis that can accurately assign a sentiment polarity score to a given opinionated text.
4. Test the developed approaches on real world data.

1.2 Dissertation Structure

- **Chapter 2** details the background and literature relating to ABSA. This chapter discusses Aspect Detection and Sentiment Analysis. Aspect Detection techniques such as Topic Modelling are discussed. Sentiment Analysis methods such as classification-based and lexicon-based approaches are discussed.
- **Chapter 3** outlines the datasets used to test our approach.
- **Chapter 4** explains the methodology for the Aspect Detection phase, as well as, addressing our test results.
- **Chapter 5** explains the methodology for the Sentiment Analysis phase, as well as, addressing our test results.
- **Chapter 6** concludes the dissertation by evaluating the dissertation overall and discusses future work.

Chapter 2: Background

Classifying opinion text at the document level can be extremely valuable. However, it does not display the whole picture instead it is representative of an overall opinion. For instance, the overall opinion of a customer about a restaurant may be negative, however, that does not necessarily mean the customer dislikes everything about the restaurant. To obtain the whole picture, we must explore the opinion at the aspect level. By using ASBA, we can discover deeper meanings behind opinions.

In this chapter, we will explore the literature and background behind the main components of ABSA. Structurally this chapter will consist of four subsections. In section 2.1 we will detail the literature relating to Aspect Detection. In section 2.2 we will look at the main features of Topic Modelling. In section 2.3 we will look at Sentiment Analysis. In the final section, we will provide a short summary of the literature.

2.1 Aspect Detection

Hu and Liu (2004) developed the first Aspect Detection approach. Their approach involved finding frequently occurring nouns and noun phrases. To obtain nouns and noun phrases they used Association Rule Mining (ARM) based on the Apriori Algorithm. They used this approach because vocabulary tends to converge when various aspects of a product are discussed. Therefore, frequently occurring nouns typically represent significant aspects. Their approach accurately detects aspects related to a single noun, however, it is inadequate when aspects are comprised of countless low-frequency terms. This approach was then improved upon by Popescu and Etzioni (2005) with their OPINE system. Their approach identifies irrelevant noun phrases, i.e. noun phrases that are not associated with product aspects. The OPINE system is 22% more precise at Aspect Detection task than Hu and Liu's approach.

Supervised approaches typically use Conditional Random Fields (CRF) [Jakob and Gurevych, 2010; Mitchell et al., 2013; Shu et al., 2017]. CRF was first proposed by Lafferty et al. (2001). CRF is a probabilistic approach for obtaining and classifying sequential data and is usually more accurate than other supervised learning algorithms, such as Support Vector Machines (SVM). CRF has proven to have high accuracy in information extraction, mainly in entity recognition (Klinger & Friedrich, 2009). Supervised approaches require pre-labelled data in order to classify future data. These approaches can be very successful, resulting in coherent aspects being discovered. However, ample human labour and expense are required to label data. Furthermore, a supervised approach is not very effective when dealing with data from different domains.

Topic modelling methods have been commonly used as an unsupervised approach to aspect detection. The need for unsupervised methods for Aspect Detection was highlighted by Titov and McDonald (2008). They revealed that comprehensive topic models such as Latent Dirichlet Allocation (LDA) [Blei et al., 2003] might not be ideal for discovering aspects. To address this, they presented multi-grain topic models to detecting local rateable aspects, where each detected aspect is a multinomial distribution over words. The benefit of this approach is that words expressing identical or similar aspects are automatically categorised under the same aspect. However, Titov and McDonald (2008) included both opinion words and aspect words in their approach, resulting in the topic model not being very accurate. Lin and He (2009) presented an ABSA model by extending LDA, however, aspect and opinion words were not clearly disconnected. Unsupervised approaches have an advantage over supervised approaches for Aspect Detection. This is mainly due to unsupervised approaches having the ability to perform well across various domains. For this dissertation we use Topic Modelling for Aspect Detection, thus we will observe Topic Modelling in more detail in the following subsection.

2.2 Topic Modelling

2.2.1 Topic modelling history

Topic modelling is a text analysis technique used to discover similarities between words in a document, where similar words are categorised together to generate topics. One of the earliest topic modelling techniques was introduced in 1997 by Landauer & Dumais where they presented the model of latent semantic analysis (LSA). Landauer & Dumais [1997] described the model as an "unsupervised high-dimensional linear associative model" that determines the connection between words in a document. This model was expanded on by Hofmann [1999] with probabilistic latent semantic indexing (PLSI), also known as probabilistic latent semantic analysis (PLSA), this was described as a text mining method built on the "spectral analysis of the term-document matrix". These ideas were further developed into the widely used and popular model of latent Dirichlet allocation (LDA) [Blei et al., 2003]. Blei and Lafferty [2007] suggested that a constraint of LDA is that it is incapable of modelling the correlation between topics, thus they developed the correlation topic model (CTM), where the topic proportions exhibit correlation via a logistic normal distribution.

2.2.2 Probabilistic topic models

In the fields of text classification and information retrieval, probabilistic topic models, such as probabilistic latent semantic analysis (PLSA) [Hofmann, 1999] and latent Dirichlet allocation (LDA) [Blei et al., 2003], have been utilised to discover useful information hidden within documents and words. Topic modelling is a statistical method that finds the probability distribution over words of a document to discover the aspects or themes within them. With topic modelling documents do not need to be annotated prior to analysis, they are unsupervised.

Probabilistic latent semantic analysis

One of the earliest examples of probabilistic topic models is the probabilistic latent semantic analysis (PLSA) which was introduced in 1999 by Hofmann. This was a milestone in topic modelling with a strong mathematical basis, a probability version of latent semantic analysis (LSA). PLSA is a statistical technique that models co-occurrence information under a probabilistic framework designed to find the core semantic structure of the data [Oneata, 2011]. With co-occurrence meaning that a matrix includes both the word and the document at the same time. The model states that words contain within them semantic information and documents that have related topics will include related words. Consequently, latent topics are found by classifying groups of words that appear together often.

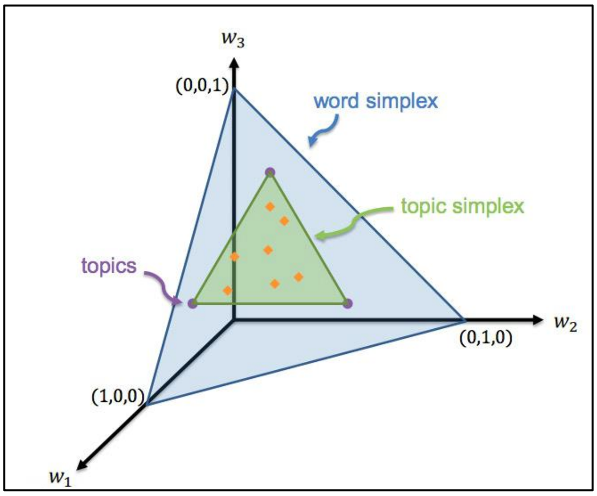


FIGURE 1: PLSA TOPIC SIMPLEX FOR THREE TOPICS
[CHEN, 2017]

The central assumption that PLSA makes is regarding the make-up of the documents, PLSA proposes that documents are multinomial probability distributions over topics, and topics are multinomial probability distributions over words. A multinomial distribution can be parameterised by a vector of real values that summarise to one. To visualise this, we imagine an individual multinomial distribution as a point on a simplex, as shown in Figure 1, where each corner of the topic simplex represents a topic/outcome. A multinomial distribution that favours a single outcome corresponds to a corner of the simplex. A multinomial distribution which favours each outcome equally corresponds to the centre of the simplex.

Advantages of PLSA:

- While computing the model PLSA employs the Expectation-maximization (EM) algorithm to solve by iteration. Thus, making it computationally faster and easier than its predecessor LSA, which utilises the singular value decomposition (SVD) algorithm.
- PLSA has well-defined probabilities and a strong statistical basis when compared with LSA.

Latent Dirichlet allocation

Blei et al. [2003] argued that PLSA is not a complete generative model since it cannot compute the probability of a new document. Also, due to the model having many parameters, the model is highly complex. This results in the model being prone to overfitting when training the data. Therefore, in 2003, Blei et al. introduced Latent Dirichlet allocation (LDA), a Bayesian interpretation of PLSA.

LDA is almost identical to PLSA except it converts the PLSA into a generative model by imposing a Dirichlet prior on the parameters of the model, resulting in the parameters being more regularized. A Dirichlet distribution is a method used to model random probability mass functions and is commonly used as a prior in Bayesian statistics. In this sense, LDA can be described as a Bayesian adaptation of PLSA. Bayesian formulation helps LDA avoid overfitting when dealing with datasets.

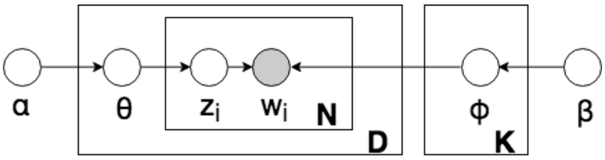


FIGURE 2: A GRAPHICAL REPRESENTATION OF LDA
IN THE MODEL ONLY WI IS SHADED, MEANING IT IS THE ONLY VARIABLE WE CAN SEE, THE REST ARE LATENT VARIABLES
[ADAPTED MODEL; BLEI ET AL., 2003]

The generative process under LDA is as follows:

1. For $k = 1 \dots K$:
   a) $\phi^{(k)} \sim Dirichlet(\beta)$
2. For each document $d \in D$:
   a) $\theta_d \sim Dirichlet(\alpha)$
   b) For each word $w_i \in d$:
      i. $z_i \sim Discrete(\theta_d)$
      ii. $w_i \sim Discrete(\phi^{(z_i)})$

LDA assumes that there is K number of topics and a fixed vocabulary (with V unique terms). Where each topic is a probability distribution $\phi^{(k)}$ over the vocabulary, and each topic has a Dirichlet prior β. If β is high, topics are expected to have a large mixture of vocabulary; if β is low, topics are expected to have a small mixture of vocabulary. LDA denotes D as the total number of documents in a corpus, and N represents the number of words within document d. θ_d represents the topic proportions for each document, and it is drawn from the Dirichlet prior parameter α. The effect of α can be seen in Figure 3, where you can see as alpha increases the topics move away from the corners of the simplex, this results in a smooth distribution. If the parameters for the Dirichlet distribution fix to 1, the LDA model will downgrade to a PLSA model. Each word in document d is represented by w_i. Where z_i is the topic distribution for w_i.

The generative process leads to the following joint distribution:

$$p(w, z, \theta, \phi | \alpha, \beta) = p(\phi | \beta) p(\theta | \alpha) p(z | \theta) p(w | \phi_z) \quad (1)$$

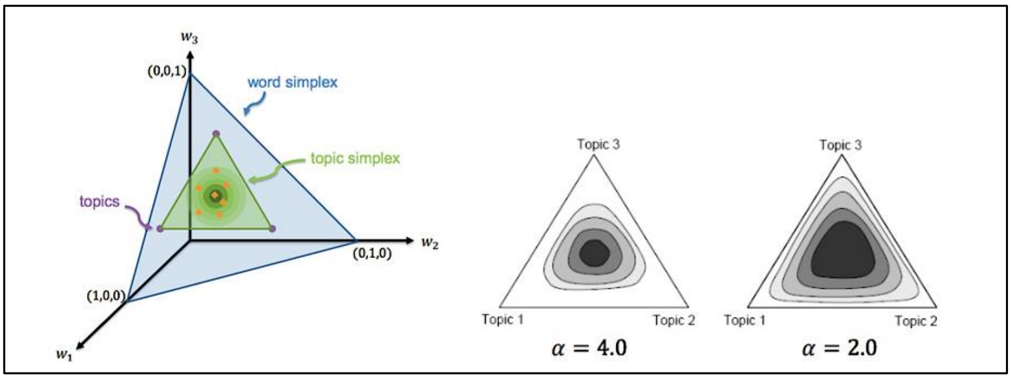


FIGURE 3: LDA TOPIC SIMPLEX FOR THREE TOPICS
[CHEN, 2017]

Using LDA, topics can be determined unsupervised by analysing the original text. Sets of words that frequently co-occur in a single document are assigned high probabilities, resulting in the creation of topics. Words that appear in the same sentence are usually semantically related to each other, meaning the topics are typically easily interpretable and can be summarised with a single word or phrase. Table 1 shows three topics that were discovered, each topic illustrates the top five most frequent words sorted in descending order by probability [Blei, 2012]. From this, one can easily determine the category associated with each topic. For instance, the words in Topic 1: *protein*, *cell* and *gene* can all be listed under the title Protein, Topic 2 could be categorised as *Tumour*, and Topic 3 could be categorised as *Computation*.

Topics	Terms
Topic 1	Protein, Cell, Gene, DNA, Polypeptide
Topic 2	Tumour, Cancer, Disease, Death, Medical
Topic 3	Computer, Model, Algorithm, Data, Mathematical

TABLE 1: LDA TOPIC MODEL TERMS EXAMPLE
[BLEI, 2012]

2.2.3 MODEL ESTIMATION

Posterior inference is the main issue in topic modelling. Posterior inference is the process of learning the posterior distributions of latent variables by observing the generative process in reverse. In LDA, this is used to solve the following equation:

$$p(\theta, \phi, z|w, \alpha, \beta) = \frac{p(\theta,\phi,z,w\,|\,\alpha,\beta)}{p(w\,|\,\alpha,\beta)} \qquad (2)$$

This distribution can be incredibly difficult to compute. However, numerous learning algorithms have been developed to solve the problem by model estimation to infer latent variables [Asuncion et al., 2009; Blei and Lafferty, 2009]. The main algorithms are as follows:

- **Maximum likelihood estimation** (MLE) [Hofmann, 1999]. MLE finds the parameter values that maximize the likelihood function for the PLSA model.
- **Expectation propagation** (EP) [Minka and Lafferty, 2002; Griffiths and Steyvers, 2004]. EP attempts to find approximations to a probability distribution.
- **Variational Bayesian inference** (VB) [Blei et al., 2003]. VB delivers an approximated posterior probability using a variational distribution.
- **Gibbs sampling** [Griffiths and Steyvers, 2004] is a sampling method in which a Markov chain is constructed to be used as a sample from a conditional distribution.
- **Collapsed variational Bayesian inference** [Teh et al., 2006a] is similar to VB, however, the variables θ and β are collapsed.

In this dissertation, we will focus on Gibbs sampling, which will be discussed further in the following subsection.

2.2.4 GIBBS SAMPLING

Gibbs sampling utilises a Markov Chain Monte Carlo (MCMC) framework [Gilks, Richardson and Spiegelhalter, 1998]. MCMC algorithms construct a Markov chain where the target distribution is the equilibrium distribution. By observing the chain over numerous iterations, a sample can be obtained. The greater the number of iterations, the more accurate the sample is to the target distribution.

In Gibbs sampling, a Markov chain is run iteratively to obtain a sample using the conditional distributions of posterior variables. The chain is run until the samples start to converge. When employed for LDA, we are concerned with topic proportions for each document θ_d, the topic-word distributions ϕ_z, and z_i is the topic distribution for each word.

2.2.5 MODEL EVALUATION

Model evaluation is an important step in the topic modelling process. The model's performance must be measured to ensure the model is accurate and efficient. For instance, deciding the number of topics k is a problem that is generally encountered if the parameter is undefined [Blei and Lafferty, 2009]. In the following subsections, we will explore the methods used to measure performance, both in terms of data and in terms of human judgement.

Performance using data

The most common metrics used to measure performance using data in topic modelling are as follows:

- **Perplexity** is the most widely used method of evaluating a probabilistic topic model. This calculates the log-likelihood of a held-out test data. Where the lower the perplexity score the better the model [Blei et al., 2003]. Chang et al., (2009) found that perplexity and human judgement are not correlated.

- **Empirical likelihood** [Li and McCallum, 2006] measures the empirical likelihood function depending on constraints established on the estimating function and the assumption that probability weights of the function sum to 1 [Mittelhammer et al., 2000].
- **Marginal likelihood** (ML) "is a central quantity in Bayesian model selection and model
  averaging. It is defined as the integral over the parameter space of the likelihood
  times the prior density" [Raftery et al., 2007]. ML can be estimated by:
  - **Harmonic mean estimator** [Griffiths and Steyvers, 2004]
  - **Chib-style estimation**, [Chib, 1995]
  - **Left-to-right samplers** [Del Moral et al., 2006]

Performance using human judgement

Chang and Blei (2009), introduced metrics intended to increase the evaluation accuracy of using human judgement for topic models. They presented two metrics:

- **Word intrusion** measures:
  - if the words in each topic semantically related
  - if the topics resemble typical human categories
- **Topic intrusion** measures:
  - If the topics represent aspects a human would generally expect from the document

Chang and Blei (2009) found, after comparing the performance of tradition measures with human judgement, that traditional measures did not result in coherent topic models. Thus, they suggest that topic models would be better evaluated using "real word task performance" rather than traditional likelihood-based measures.

2.3 Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a computational technique used to capture a writer's opinions and attitudes, towards an issue, conveyed in text (Liu and Zhang, 2012). Opinions are usually expressed as positive, negative, or neutral. Sentiment analysis is being widely used by companies to understand customer opinions expressed in the form of tweets, reviews, etc.

Sentiment analysis generally attempts to capture the opinion of a writer or speaker in toward a specific topic or item. With enough opinions mined a researcher draw significant insights regarding customer attitude, this is usually applied to marketing or product development.

There are two main approaches to applying sentiment analysis: Classification-based Approach, and Lexicon-based Approach. These will be discussed further in the following subsections.

2.3.1 Classification-based approach

Sentiment analysis can be practised using classification based supervised learning techniques. Supervised learning is the machine learning task whereby using given data $(x_1, y_1), \dots, (x_n, y_n)$ (where x_i denotes the data points, and y_i denotes the class/value) a function is obtained in the form of $g: X \to Y$ (where X denotes the input space, and Y denotes the output space). Supervised learning obtains a function from training data (a subset of the data). Classification based supervised learning is used when the y_i are from a finite set (i.e. not real-values).

Sentiment analysis can be applied as a supervised learning classification problem. For instance, imagine we have a dataset of product reviews that have been assigned a rating between 1 and 5 stars, the review text is the input and the output is the star rating which is from a finite set of numbers. The dataset can be split into training and test data; hence, a function can be obtained using the training data. Applying the function on the test data would retrieve an accuracy rate. If the function is accurate it could be used to predict the rating of consequent reviews.

Many supervised learning techniques can be used for sentiment analysis. For instance, Pang et al. (2002) used support vector machines (SVM) and naïve Bayes to classify movie reviews as either positive or negative. They concluded that naïve Bayes and SVM were accurate when using unigrams as features. Further research, by Pang and Lee (2008), led to the discovery of additional features. The main features are as follows:

- Opinion words and phrases: Opinion words are words that are generally used to convey sentiments. Such as “good” and “bad”, or “beautiful” and “ugly”. Opinion phrases are idioms commonly used to convey sentiments. Such as the phrase “to pay through the nose” which often denotes paying an excessive amount for something.
- Negations: Negations often alter the orientation of an opinion. For instance, the sentence “It is not good” conveys a negative sentiment. The word “not” alters the opinion of the sentence from positive to negative.
- Term frequency: This includes the frequency count with the individual word.
- Part-of-speech (POS) tagging: Adjectives were found to convey opinions.

Researchers have also developed alternative techniques for sentiment classification. For instance, Dave et.al, (2003) developed a score function used to convey positive and negative words from reviews. Classification accuracy was also improved by employing weighting schemes, [Paltoglou and Thelwall, 2010]. Tan et al. (2008) increased the efficiency of labelling training data by labelling a sample of revealing instances which are then used to train a supervised classifier. Melville et al. (2009) sought to increase sentiment accuracy by including the lexical knowledge of opinion words.

There are some fundamental problems with using Classification-based approaches for sentiment analysis. The main being that Classification based methods are only consistently accurate when trained on a corpus of a specific domain. Thus, this approach may be unsuitable for customer reviews, where products and services come from various domains.

2.3.2 Lexicon-Based Approach

A Lexicon-based approach uses a list of words that convey a sentiment such as good, bad, happy, and sad. These could also include opinion phrases, as described in the previous section. These are known collectively as an Opinion Lexicon. An opinion lexicon can be utilised to improve sentiment classification and can be represented in different ways. For instance, an opinion lexicon can be represented as:

- A binary classification of positive and negative words.
- A range of sentiment polarity grades, such as excellent, good, neutral, poor, terrible.
- A real value, such as a binary measure of 1 and -1, or a range from 1 to 5.

An opinion lexicon is generally compiled using three main approaches. These are the dictionary-based approach, the corpus-based approach, and the manual approach. The manual approach is inefficient and is rarely used. Thus, we will be discussing the dictionary-based approach and the corpus-based approach.

Dictionary-based Approach

The dictionary-based approach is a widely used approach, where usually a small list of opinions words is expanded using common synonyms and antonyms typically found by using an online dictionary like WordNet [Miller et al., 1990]. The small list of opinion words is manually classified according to their orientation (positive, or negative). Hu and Liu (2004), and Kim and Hovy

(2004) used this method. This approach does, however, not consider domain specific opinions. For instance, if a boombox is described as loud, it is positive.

Corpus-based Approach

The corpus-based approach also uses a list of opinion words; however, it also utilises syntactic and co-occurrence patterns to discover opinion words in a corpus. For instance, Hazivassiloglou and McKeown (1997) used a list of manually annotated adjectives based on their semantic orientation and calculated a log-likelihood ratio for all to words depending on the frequency at which the words co-occur in the same sentence. The words were then clustered into two sets of words: positive, and negative.

The issue, however, with using a corpus-based approach is that does not usually identify all the opinion words in the English language. Whereas a dictionary approach does. Although, a corpus-based approach does have its advantages over a dictionary-based approach. The main benefit is a corpus can find domain-specific words, such as slang, common spelling errors, and context-specific words.

2.4 SUMMARY

For Aspect Detection, we found that both of the two main approaches (supervised, and unsupervised) have significant benefits and flaws. Supervised approaches result in coherent aspect, but are extremely domain specific. Unsupervised methods are not domain specific but can result in incoherent results. Researchers have attempted to rectify unsupervised methods by extending topic modelling, but have not succeeded in separating opinion words from aspect words [Titov and McDonald, 2008; Lin and He, 2009].

For Sentiment Analysis, we found that the common methods used are also significantly flawed. Classification-based approaches although accurate, are not very effective when used with different domains. The two Lexicon-based approaches (dictionary-based, and corpus-based) also have notable disadvantages. A dictionary-based approach will be unable to correctly identify context-specific words, whereas a corpus-based approach will likely result in a limited vocabulary.

CHAPTER 3: DATA

For this dissertation we collected three separate datasets, all the datasets are of customer reviews. The purpose of the three datasets is to test our approach. Datasets 1 & 2 are used to test our Aspect Detection approach, as well as apply our Sentiment Analysis approach on; data from two different domains was collected in order to demonstrate domain independence. Dataset 3, were collected for the sole purpose of building an Opinion Lexicon to use for the Sentiment Analysis. The three datasets are addressed in the following subsections. Since we are not collecting data containing personal information ethical approval is not needed.

3.1 DATASET 1 – HOTEL REVIEWS

Dataset 1 is the primary dataset we used to test our approach. The dataset is a collection of TripAdvisor reviews, collected from Ganesan and Zhai, (2011). The dataset consists of full reviews of hotels from 10 different cities, with a total of 259,000 reviews. Each review contains the review title and full review.

3.2 DATASET 2: AMAZON REVIEWS

Dataset 2 was collected in order to demonstrate domain independence. The dataset is a small collection of Amazon customer reviews. This dataset was scraped from Amazon.com, using the R packages "pacman" [Rinker and Kurkiewicz, 2017]and "rvest" [Wickham, 2016]. The dataset has a total of 35,560 reviews for 11 different products. The dataset consists of two variables, the stars (star rating), and the comments Each comment has a corresponding star rating ranging from 1 to 5.

3.3 DATASET 3: REVIEWS COLLECTION

Dataset 3 was collected for the purpose of building an Opinion Lexicon to use for the Sentiment Analysis. We obtained an Amazon review dataset from the Stanford Network Analysis containing 3,650,000 reviews spanning May

1996 - July 2014 [McAuley and Leskovec, 2018]. The dataset has reviews of 328,245 unique products. The dataset consists of three variables, the stars (star rating) which, the review comments, and the review titles. Each review has a corresponding star rating ranging from 1 to 5.

3.4 Testing Software

The software R version 3.5.1 run on RStudio Version 1.0.153 will be used to test our Model [R Core Team, 2017].

CHAPTER 4: ASPECT DETECTION

Our main objective for Aspect Detection is to present a versatile and scalable approach that discovers coherent aspects. There are two main approaches to Aspect Detection – supervised and unsupervised.

Supervised Aspect Detection requires pre-labelled data in order to classify future data. This approach can be very effective, resulting in coherent aspects. However, ample human labour and expense are required to label data, whereas unlabelled data is usually readily available. Additionally, a supervised approach is not usually robust enough to deal with the vast range of products and services being reviewed, the model would lack the versatility to perform well across domains. Supervised methods also struggle with the nature of online customer reviews. Online customer reviews typically contain spelling errors, grammatical errors, slang, and jargon. This could prove problematic for supervised methods that rely on dictionaries, resulting in important aspects being omitted.

Unsupervised Aspect Detection does not require labelled data and instead relies mainly on statistics to classify data. Since unsupervised approaches do not require labelled data, they are generally robust and can easily be transferred across domains and languages. Topic Modelling is a frequently used unsupervised method for Aspect Detection (Mei et al., 2007; Titov and McDonald, 2008; Lin and He, 2009; Moghaddam and Ester, 2011). However, previous methods have not adequately addressed the coherence issue associated with Topic Models and unsupervised approaches in general. Mimno et al. (2011) suggested that the problem with Topic Modelling is that it generates 'junk' topics. This is because unsupervised models generally conflate semantics with words. An unsupervised model does not consider the meaning of the word, and so, it can be ineffective when trying to extract precise aspects from documents. However due to unsupervised methods, not considering semantics, they perform well with spelling errors, grammar, etc.

This chapter will be structured into six parts. In the first section, we will present our NOA-LDA system. In section 4.2, we will address our pre-processing approach. In section 4.3, we will describe the implementation used to conduct the LDA. In section 4.4 we reveal our Aspect Detection experiments and test results. In the final section, 4.5, we will discuss our results and approach.

4.1 NOA-LDA, A Noun-only approach to Latent Dirichlet Allocation

For the Aspect Detection, we propose that an unsupervised approach is most desirable. This is because customer reviews are being produced for a variety of products and services, so the approach needs to be robust and versatile with the ability to perform across domains and languages. To address the coherence problem associated with an unsupervised approach, we propose a Noun-Only Approach (NOA) to the topic modelling method LDA, resulting in our NOA-LDA system. The reason for using LDA as the chosen topic model is that it is a well-researched method with a stronger mathematical basis than similar topic models, such as LSA [Landauer & Dumais, 1997] and PLSA [Hofmann, 1999]. Hu and Liu (2004), found that nouns typically represent aspects in product reviews. Thus, by conducting LDA using solely nouns, we can solve the coherence problem associated with Topic Models. This approach will allow us to retain the versatility and robustness of LDA. LDA will be used in conjunction with Part-Of-Speech (POS) Tagging which will feed the LDA only noun words. The POS tagger is pre-trained; thus, no manual labelling is needed. The POS tagger also works across many languages, making it reasonably versatile.

4.2 NOA-LDA Text Pre-processing Approach

Adequate pre-processing is necessary to ensure the LDA model results in coherent aspects. Figure 4 gives an outline of the NOA-LDA pre-processing method for Aspect Detection.

```
Model: NOA-LDA Text Pre-Processing for Aspect Detection
Input: Review Text
Method:
FOR each document $d \in D$
    Extract Sentences $s \in S$ from each document $d$
    FOR each sentence $s$ in document $d$
        Perform tokenisation
        FOR each word $w_i$ in sentence $s$
            Normalise
                Set all characters to lower case
                Remove numbers
                Remove punctuation
                Strip whitespace
                Remove stop-words
                Remove name of the product/service and brand/company
                Use lemmatisation
            Use POS tagging to extract nouns
END FOR
Output: Noun only Review Text
```

Figure 4: NOA-LDA Aspect Detection Text Pre-processing outline

The first step in preparing the data is to extract sentences $s \in S$ from each document (comment) d. The reason for this is because each sentence usually conveys an opinion regarding a single aspect. For example, have a look at the following comment taken from Dataset 2:

> *"The noise elimination is the best I've used and working in a noisy office that is critical. They are also comfortable to wear all day. I took off one star because of the wired option, it's a non-standard connection."*

In the comment, there are three sentences. The first sentence is regarding noise cancellation, second is comfort, and the final sentence discusses connectivity. This is a common pattern amongst most product reviews, where a variety of aspects are discussed in a single review. Therefore, it is difficult to assign an entire review a single aspect without causing conflict. However, assigning each sentence with an aspect is much simpler.

The next step is to split the text further, i.e. we tokenise the text. Tokenisation is the process of splitting strings into smaller pieces. In this case, we want to split the sentences into words w_i, this will make the data ready for further processing. The text must then be normalised. This is the process of converting all the text to the same format. To do this we perform a series of subtasks:

- Set all characters to lower case
- Remove numbers
- Remove punctuation
- Strip whitespace
- Removing stop-words. Stop-words are words that convey little or no meaning, such as, “and”, “the”, and “a”. Removal of these words will likely increase the coherence of the aspects.
- Removing the name of the product/service and brand/company from the corpus. This will improve the coherence of the model since the name of the product/service and brand/company occur frequently and are not aspects.
- Lemmatisation. Lau et al. (2014) found that lemmatisation improved aspect coherence. Lemmatisation is the process of normalising (returning) words into their dictionary form or base; this is known as the lemma. For instance, “houses” becomes “house”, or “running” becomes “run”. Morphological analysis can also remove stop-words.

The data is then annotated with POS tagging. This is performed in order to extract only nouns from the data. The POS tagging and lemmatisation is performed using the R package “UDPipe” [Wijffels, 2018].

4.3 LDA Implementation

For the LDA we use a standard implementation method. Figure 5 gives an outline of the LDA method used for Aspect Detection. Terms in Figure 5 are detailed in Section 2.2.2.

```
Model: LDA for Aspect Detection
Input: Noun Only Review Text
Method:
FOR each aspect $k = 1 \dots K$
        Draw a multinomial word distribution $\phi^{(k)} \sim Dirichlet(\beta)$
END FOR
FOR each document $d \in D$
        Draw an aspect distribution $\theta_d \sim Dirichlet(\alpha)$
        FOR each sentence $s$ in document d
        FOR each word  $w_i$ in sentence $s$
                Draw an aspect $z_i \sim Discrete(\theta_d)$
                Draw a word $w_i \sim Discrete(\phi^{(z_i)})$
END FOR
Output: Aspect label assignment for all Sentences in the Review Text
```

Figure 5: LDA outline

4.3.1 Selecting the number of aspects

The first step in building the LDA model is to determine the number of topics (aspects in our case) (k). There are many techniques for doing this, one method is by using human judgement through trial and error. However, this can be very time consuming since you have no idea as to how many aspects you are looking for. A simple solution would be to follow a data approach to predicting k, this will give us an estimate for the ideal number of k. After we have an estimate for the number of aspects, a trial and error approach could be used. The data approach for estimating k we will be using is the harmonic mean estimator [Newton and Raftery (1994); Griffiths and Steyvers (2004); Griffiths et al. (2005); Wallach (2006)]. The method has been chosen due to its efficiency and ease of use. The method computes the log-likelihood of $p(w|\ k)$, where w is the words in the corpus. To implement this method, we will use the R package "topicmodels" [Grün and Hornik, 2011]. Once we have the number of aspects, we need to decide which algorithm we will be using for the model estimation.

4.3.2 Model estimation and evaluation

The algorithm used for the model estimation is the Gibbs sampler [Griffiths and Steyvers, 2004]. The main benefit of using the Gibbs sampler is tuning is not essential for the Markov chain to converge. The sampler is also easy to apply and is computationally efficient. In order to increase the performance of the model, some parameters can be set. The Dirichlet parameters α and η can be specified. Griffiths and Steyvers (2004) propose a value of $50/k$ for α and 0.1 for η, we follow these settings in our implementation. The default setting for the number of Gibbs sampling iterations (4000) was used. After we specify all these variables we can run the LDA model. The LDA model will be implemented using the same R package mentioned above "topicmodels", this package allows the user to select Gibbs sampling and the parameters.

To evaluate the results, we will use human judgement. We will examine the top 5 terms per topic discovered to determine whether the words that have been allocated together are in fact related. If so, the topics are labelled according to the aspect they most relate to. This data is then joined to the comments, ready for the sentiment analysis.

4.4 ASPECT RESULTS

To demonstrate our Aspect Detection approach, we used two examples:

- Hotel reviews: Reviews relating to a Hotel from Dataset 1
- Headphone reviews: Reviews relating to Noise Cancelling Headphones from Dataset 2

This allows us to test the method on two different domains.

The data was first split into sentence level using the R package "lexRankr" [Spannbauer and White, 2017]. Resulting in 21,253 sentences. The data was then tokenised and normalised using the R packages "dplyr" [Wickham et al., 2017] and "tm" [Feinerer and Hornik, 2018]. The name of the product/service and the brand/company were removed from the reviews. To test the accuracy our NOA, we performed the LDA twice for each example, with one corpus containing only nouns, and with one corpus containing all parts-of-speech; allowing us to compare the results.

The first step in building the LDA model was to discover the number of aspects (k). Using the harmonic mean estimator, we found that the ideal number of aspects for the Hotel reviews with all POS was 10. This is shown in Figure 6; the highest point of the log-likelihood represents the ideal number of aspects. The ideal number of aspects for the Hotel reviews with only nouns was 4, this can be seen in Figure 7. The ideal number of aspects for the Headphone reviews with all POS was 10, this can be seen in Figure 8. The ideal number of aspects for the Headphone reviews with only nouns was 4, this can be seen in Figure 9.

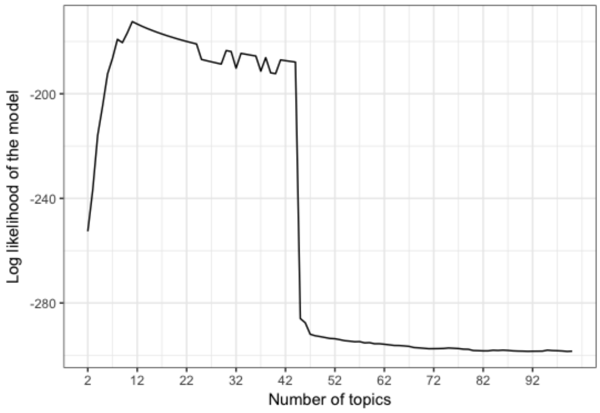


FIGURE 6: HOTEL REVIEWS – IDEAL NUMBER OF ASPECTS FOR ALL POS

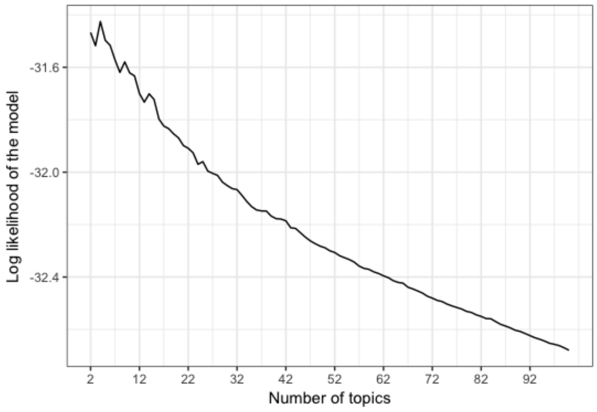


FIGURE 7: HOTEL REVIEWS – IDEAL NUMBER OF ASPECTS FOR ONLY NOUNS

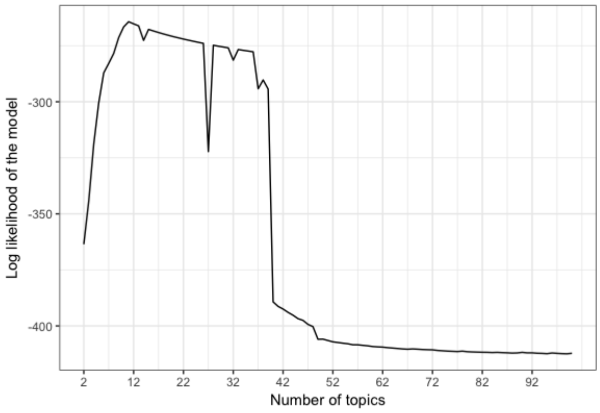


FIGURE 8: HEADPHONE REVIEWS – IDEAL NUMBER OF ASPECTS FOR ALL POS

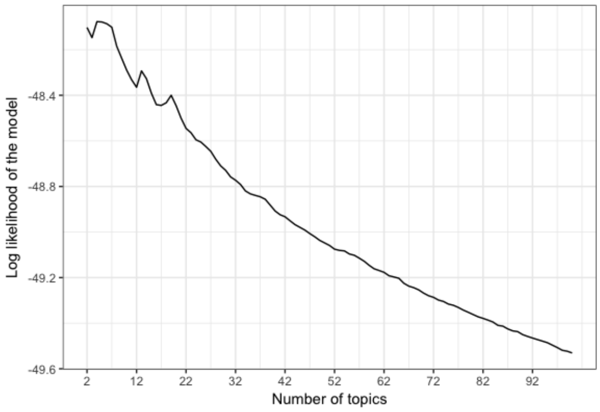


FIGURE 9: HEADPHONE REVIEWS – IDEAL NUMBER OF ASPECTS FOR ONLY NOUNS

We then ran the LDA with Gibbs sampling on both examples. The time taken to run the LDA Model on each example and variation is shown in table 2. To evaluate the results, we found the top 5 terms per Aspect, where the terms are in descending order with respect to importance (Table 4; Table 4; Table 5; Table 6).

Data	LDA Run Time (seconds)
Hotel reviews – All POS	388.389
Hotel reviews – Only Nouns	40.823
Headphone reviews – All POS	752.633
Headphone reviews – Only Nouns	48.158

TABLE 2: LDA RUN TIME

Aspects	Terms
Aspect 1	staff, friendly, helpful, check, nice
Aspect 2	time, thank dear, comments, taking
Aspect 3	stay, review, thank, dear, experience
Aspect 4	room, clean, small, bed, comfortable
Aspect 5	stayed, one, night, booked, nights
Aspect 6	great, also, team, can, read
Aspect 7	location, station, close, walk, just
Aspect 8	hear, stay, sorry, like, enjoyed
Aspect 9	good, breakfast, service, excellent, bar
Aspect 10	rooms, however, first, refurbishment, air

TABLE 3: HOTEL REVIEWS – TOP 5 TERMS FOR ALL POS

Aspects	Terms
Aspect 1	staff, service, breakfast, hear, team
Aspect 2	location, station, night, king, train
Aspect 3	stay, experience, advisor, thank, guest
Aspect 4	room, size, bed, bathroom, air

TABLE 4: HOTEL REVIEWS – TOP 5 TERMS FOR ONLY NOUNS

Aspects	Terms
Aspect 1	sound, quality, good, great, amazing
Aspect 2	battery, time, life, long
Aspect 3	can, music, hear, without, still
Aspect 4	pair, best, ive, used, now
Aspect 5	bluetooth, phone, audio, devices, device
Aspect 6	use, work, even, well, lot
Aspect 7	just, like, get, dont, really
Aspect 8	better, wireless, much, price, worth
Aspect 9	will, one, buy, new, issue
Aspect 10	comfortable, ear, also, ears, head
Aspect 11	noise, cancelling, cancellation, great, well

TABLE 5: HEADPHONE REVIEWS – TOP 5 TERMS FOR ALL POS

Aspects	Terms
Aspect 1	bluetooth, device, phone, call, connection
Aspect 2	life, battery, issue, problem, charge
Aspect 3	noise, cancellation, cancel, reduction, cancelling
Aspect 4	sound, music, bass, volume, cord

TABLE 6: HEADPHONE REVIEWS – TOP 5 TERMS FOR ONLY NOUNS

4.5 DISCUSSION

Our main objective for Aspect Detection was to present a versatile and scalable approach that discovers coherent aspects.

In terms of the scalability, run time for the LDA model was significantly faster using the NOA-LDA system. In Table 2, we can see that the LDA runtime for all POS was significantly longer than the NOA-LDA in both examples. With the Hotel reviews performing approximately twice as fast with a NOA-LDA, and the Headphone reviews performing approximately 15 times faster with an NOA-LDA.

In terms of coherence, our approach also succeeded. If we look at Table 3, we can see the top 5 terms per aspect for the Hotel review corpus with all POS. The majority of these aspects were incoherent. Only 3 out of the 10 aspects can be strongly related to a single aspect; aspect 1 described as staff friendliness, aspect 4 can be described as room quality, and aspect 7 as location. The remaining aspects cannot be inferred. Whereas, in Table 4, the top 5 terms per aspect for the Hotel review corpus with only nouns was considerably more coherent. All 4 of the aspects discovered were coherent; aspect 1 can be described as service quality, aspect 2 can be described as Location, aspect 3 can be inferred as stay experience, and aspect 4 can be described as room quality. Thus, the inferred aspects are as follows:

- Aspect 1: Service Quality
- Aspect 2: Location
- Aspect 3: Stay Experience
- Aspect 4: Room Quality

From the LDA results, it is evident that using a NOA-LDA produces more accurate aspects. The NOA-LDA system is also significantly more scalable. Additionally, the approach is versatile, demonstrating its ability to infer coherent aspects across two very different domains.

Coherence also significantly improved for the Headphone reviews. If we look at Table 5, we can see the top 5 terms per aspect for the Headphone reviews corpus with all POS. Much of these aspects were incoherent. Only 4 out of the eleven aspects can be strongly related to a single aspect; aspect 1 described as sound quality, aspect 2 can be described as battery life, aspect 5 as connectivity, and aspect 11 as noise cancellation. The remaining aspects cannot be inferred. Whereas, in Table 6, the top 5 terms per aspect for the Headphone reviews corpus with only nouns was significantly more coherent. All 4 of the aspects discovered were coherent and could be attributed to a single aspect. Aspect 1, can be described as connectivity. In aspect 2 the words life, battery, and charge all relate to battery life. All the words in aspect 3 relate to noise cancellation. The words cancel and cancelling are not commonly used as nouns, yet they are classed as nouns. This could be because of the way the words are used within the context of the sentence. Aspect 3 can be described as sound quality, since the words sound, music, bass, and volume all relate to sound. Thus, the inferred aspects are as follows:

- Aspect 1: Connectivity
- Aspect 2: Battery Life
- Aspect 3: Noise Cancellation
- Aspect 4: Sound Quality

Overall our Aspect Detection approach met our objectives of presenting a versatile and scalable approach that discovers coherent aspects.

Chapter 5: Sentiment Analysis

Our main objective for the Sentiment Analysis is to present a versatile and pragmatic approach. There are two main approaches to Sentiment Analysis – the Classification-based approach, and the Lexicon-based approach.

The classification-based approaches for Sentiment Analysis are very popular amongst researchers, however, such approaches are not ideal for Sentiment Analysis. This is because Sentiment Analysis and Text Classification are considerably different. Text Classification attempts to classify new data using predefined classes. Whereas, Sentiment Analysis attempt to measure the opinion polarity of the input text. Thus, the Machine-Learning techniques used for Text Classification are not ideal for Sentiment Analysis.

Using a Classification-based approach is accurate when working with trained data from a single domain to predict new data from the same domain. For instance, if we use a Classification-based approach to train a dataset of horror movie reviews that have been pre-labelled with sentiment scores, the classifier will be able to accurately predict the sentiment of new horror movie reviews. However, if you feed the classifier reviews of family movies it will not be as accurate; this is because words such as 'scary' will likely indicate a positive sentiment for horror movies, whereas the opposite is true for family movies. This is the fundamental problem with using Classification-based methods for Sentiment Analysis, they are extremely domain specific.

Since we want to produce a method that will be accurate across a variety of products and service, a lexicon-based approach is more suitable. There are two main techniques for a lexicon-based approach: a dictionary approach, and a corpus approach.

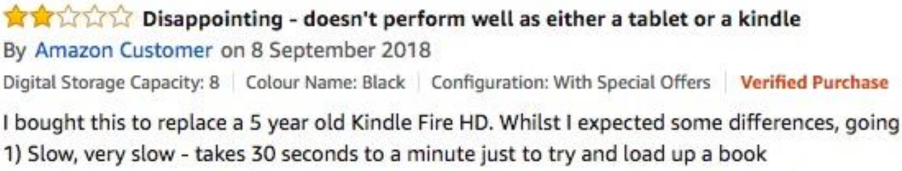
The dictionary approach uses a seed list of opinion words and expands the list using synonyms. The problem with using a dictionary approach is human judgement. The dictionary approach requires manual sentiment classification of words, this is rather simple when words are classified with a binary sentiment polarity measure (i.e. positive/negative, 1/-1). However, customer reviews are usually measured with a larger scale, such as 1-5, or 1-10, the result of manual scoring is more subjective. The corpus approach has the objective of providing a dictionary associated with a particular domain, and so is more accurate when used for that particular domain. However, producing a comprehensive list containing all English words is difficult using a corpus approach.

For our method we propose a novel Lexicon-based approach that draws insights from both the dictionary and corpus approach. Our approach is discussed in the following subsections. In section 5.1 we present our approach and the intuition behind it. In section 5.2 we will detail our Lexicon construction approach. In 5.3 we will detail how the Lexicon will be applied. In 5.4 we will look at the test results of using our approach. In the final section, 5.5, we will discuss the test results and the effectiveness of our approach.

5.1 PLSS, Pragmatic Lexicon Scoring System

The problem with the dictionary-based lexicon approach is the subjective nature of polarity scoring as mentioned previously. Our assumption is that manually scored Lexicons are subjective and are not representative of how customers score products and services in reality. Thus, we present PLSS a Pragmatic Lexicon Scoring System. The main objective of the system is to develop a Lexicon list with realistic polarity scores, which are a true representation of customer scoring attitudes. To achieve this objective the system uses a collection of reviews to extract adjectives (using POS tagging) from the title of each review, along with the rating associated with the review the adjective belonged to. The intuition behind this is the assumption that adjectives used in review titles typically convey the overall customer sentiment towards the product or service, allowing us to retrieve a sentiment score for each adjective. Whereas, full comments typically convey multiple sentiments regarding different aspects of the product, meaning the adjectives used in the full comments are less representative of the overall score.

For example, in Figure 10, a customer states that a product is "Disappointing – doesn't perform well as either a tablet or a kindle" in their title and gives a rating of two stars. In this example, our assumption is that this customer believes that the adjective "disappointing" conveys two stars. Thus, our method matches the adjectives used in the review title with the rating of the review. The construction of the PLSS is detailed in the following subsection.



★★☆☆☆ **Disappointing - doesn't perform well as either a tablet or a kindle**

By Amazon Customer on 8 September 2018

Digital Storage Capacity: 8 | Colour Name: Black | Configuration: With Special Offers | Verified Purchase

I bought this to replace a 5 year old Kindle Fire HD. Whilst I expected some differences, going

1) Slow, very slow - takes 30 seconds to a minute just to try and load up a book

FIGURE 10 – EXAMPLE AMAZON REVIEW

5.2 PLSS CONSTRUCTION

Our system uses a large collection of pre-labelled reviews to extract adjectives and their associated ratings. The adjectives will be used to build a seed lexicon list which will be expanded using synonyms. An outline of our PLSS construction can be seen in Figure 11.

```
Model: PLSS Construction
Input: Text Reviews
Method:
FOR each document $d \in D$
        Perform tokenisation
        FOR each word $w_i$ in document $d$
                Normalise
                        Set all characters to lower case
                        Remove numbers
                        Remove punctuation
                        Strip whitespace
                Use POS tagging to extract adjectives $a \in A$
                FOR each adjective $a$ in document $d$
                        Calculate the mean rating
                        Expand using synonyms
END FOR
Output: Opinion Lexicon List
```

FIGURE 11: LEXICON CONSTRUCTION APPROACH OUTLINE

Our approach to constructing the PLSS is detailed in the following five steps:

1. The first step is to retrieve a large collection of pre-labelled (with ratings) customer reviews that cover most product and service categories. For our test we retrieved a collection of 3.65 million Amazon reviews of thousands of different products from various categories (Dataset 3), however, we were unable to retrieve a dataset with ratings that also covers services.

2. The second step of our approach is to tokenise and normalise the data for each document (review title) $d \in D$. To normalise this, we perform a series of subtasks for each word w_i in document d:
   - Set all characters to lower case
   - Remove numbers
   - Remove punctuation
   - Strip whitespace

3. Using the collection of reviews, we extract adjectives (using POS tagging) $a \in A$ from the title of each review $d \in D$, along with the rating associated with the review the adjective belonged to.

4. The fourth step is to find the mean rating of each adjective a. This solves the subjectivity problem associated with the dictionary method, the method provides a more accurate representation of how customers convey emotion. For instance, if we were to manually label the word “great” to a star rating from 1 (most negative) to 5 (most positive), we might choose 4. The problem is that this is a subjective opinion. However, if 10,000 people label the word “great” with an average of 4 stars, the result is more objective and impartial.

5. The final step of our constructing the Lexicon is to expand the list with common synonyms by using an online dictionary like WordNet [Miller et al., 1990]. This will allow us to grow our list to cover the entire English dictionary.

5.3 SENTIMENT CLASSIFICATION USING PLSS

Using the PLSS constructed Lexicon list, the next step is to assign a sentiment polarity rating to aspect labelled sentences. An outline of our Sentiment Classification approach using our Lexicon list is shown in Figure 12.

```
Model: Sentiment Classification
Input: Aspect labelled Sentences
Method:
FOR each sentence $s \in S$
        Perform tokenisation
        FOR each word $w_i$ in sentence $s$
                Normalise
                        Set all characters to lower case
                        Remove numbers
                        Remove punctuation
                        Strip whitespace
                        Remove stop-words
                Inner join words with the opinion lexicon list
END FOR
FOR each aspect $k$ in sentence $s$
        Calculate average rating
END FOR

Output: Aspect Ratings
```

FIGURE 12: SENTIMENT CLASSIFICATION OUTLINE

Our approach to applying the Lexicon is detailed below.

1. The first step is to attach the original sentences (prior to noun extraction) to the discovered aspects.
2. The second step is to split the aspect labelled sentences $s \in S$ into tokens/words w_i. Resulting in one word per row.
3. We then normalise the text.
4. The fourth step is to perform an inner-join with the Sentiment Lexicon list. This will result in each sentence being allocated a rating.

5. The final step is to group the data by the aspect and summarise with the mean rating for each aspect k in sentence s.

To do this we use the general Lexicon approach using tidy data principles, shown in Figure 13, in R [Silge and Robinson, 2016].

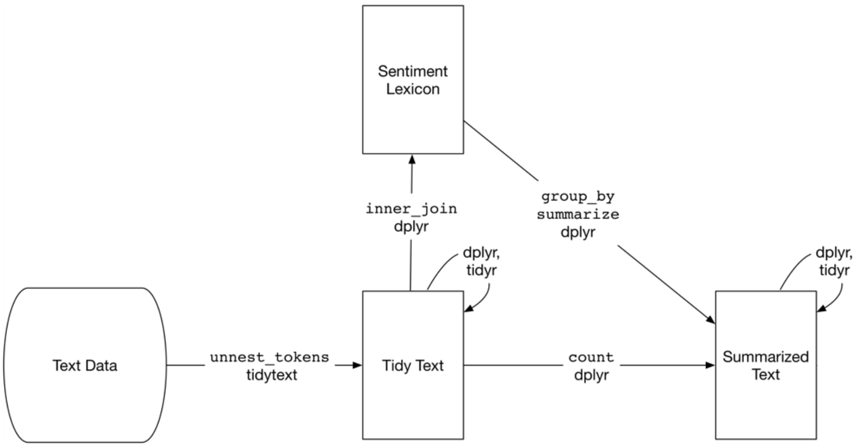


FIGURE 13 – TIDY DATA APPROACH
[SILGE AND ROBINSON, 2016]

5.4 RESULTS

We used Dataset 3 (collection of Amazon Reviews) to produce a Lexicon list using the PLSS. This resulted in 23,718 unique adjectives, along with ratings for each adjective ranging from 1 (most negative) and 5 (most positive). The data was then normalised (set all characters to lower case, remove numbers, remove punctuation, strip whitespace, remove stop-words) using the R package “tm” [Feinerer and Hornik, 2018]. The adjectives were extracted using POS tagging performed using the R package “UDPipe” [Wijffels, 2018]. However, the Lexicon could not be expanded with synonyms due to insufficient processing power. Regardless, the data produced enough adjectives to be able to proceed. To demonstrate our approach, we assigned ratings to the aspect labelled sentences from the only nouns data for the Hotel review, and the Headphone review produced in Chapter 4.3, by joining the opinion Lexicon list produced from Dataset 3. To do this we used the R package “tidytext” [Silge and Robinson, 2016]. Table 8 shows the average rating per inferred aspect for the Hotel review example. Table 9 shows the average rating per inferred aspect for the Headphone review example.

Inferred Aspects	Rating
Service Quality	3.09
Location	2.99
Stay Experience	3.04
Room Quality	3.10
AVG	3.06

TABLE 7 – HOTEL REVIEW RATINGS

Inferred Aspects	Rating
Connectivity	3.09
Battery Life	3.05
Noise Cancellation	3.22
Sound Quality	3.16
AVG	3.12

TABLE 8 – HEADPHONE REVIEWS RATINGS

5.5 DISCUSSION

Our main objective for Sentiment Analysis was to present a versatile and pragmatic approach. Our PLSS system is versatile within the domain of customer reviews, which is the basis of this dissertation. For the testing, however, we were limited to an Amazon dataset to construct a Lexicon. This only covers products, in order to fully cover the domain of customer reviews, this must be joined with a dataset that includes services.

There is no quantitative technique for measuring the accuracy of the PLSS. This is because our data was split into aspects, thus, previously allocated ratings are no longer associated. Our assumption is that manually scored Lexicons are subjective and are not representative of how customers rate products. One way to justify the pragmatism of our approach to compare it with an existing lexicon list, such as the AFINN lexicon [Nielsen, 2011]. The AFINN lexicon contains 2476 seed opinion words labelled with a rating ranging from -5 to 5, with -5 being most negative and 5 being most positive. These ratings can be adjusted to correspond with the common customer review rating system (-5 to 1, -4 to 1.4, …, 4 to 4.6, 5 to 5). Upon examining the words and their rating we realise the allocated ratings do not resemble reviewer attitudes. For instance, in the AFINN lexicon, the reviewer needs to use words such as "breathtaking" and "outstanding" in order to allocate 5 stars. This sort of vocabulary is common amongst professional critics, but not amongst the majority of customer reviews. Words such as "great" are usually enough to allocate 5 stars on Amazon, however, according to AFINN "great" is only worth 4.2. The distribution of ratings using AFINN can be seen in Figure 14, this shows that most words are rated between 2 and 3. We compare this to the distribution of Amazon review ratings (Dataset 3) shown in Figure 15. From Figure 10 we can see that Amazon review ratings are left-skewed, with most ratings being between 4 and 5 stars, and the least between 2 and 3. Thus, our PLSS replicates realistic customer rating attitudes.

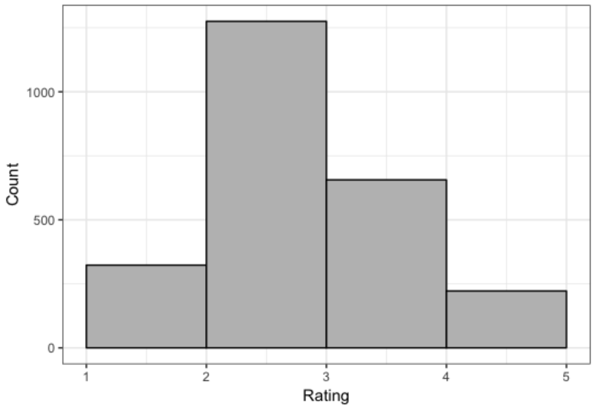


FIGURE 14: AFINN LEXICON RATING DISTRIBUTION

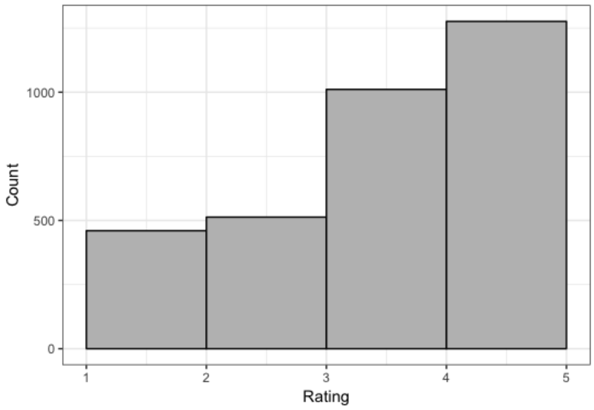


FIGURE 15: AMAZON.COM RATING DISTRIBUTION

CHAPTER 6: CONCLUSION AND FUTURE WORK

The main aim of this dissertation was to produce a novel ABSA system for customer reviews that is scalable, versatile, and that produces coherent results. In this dissertation, we have presented an ASBA system which successfully meets the criteria for our aim. In the following subsections, we will discuss our contributions as well as future work.

6.1 CONTRIBUTIONS

6.1.1 NOA-LDA

By examining relevant literature, it was determined that previous Aspect Detection methods were lacking in either coherence or versatility. Thus, we proposed a NOA to the topic modelling technique LDA (NOA-LDA) for Aspect Detection. We showed through our testing that a NOA-LDA system is a superior method compared with a raw corpus LDA for Aspect Detection; it produced more accurate and coherent aspects. This was demonstrated with two examples: Hotel reviews, and Headphone reviews. Where both examples had considerably greater coherence with a NOA-LDA system than LDA with all POS. This confirms the work by Hu and Liu (2004) whose Aspect Detection approach found frequently occurring nouns and noun phrases, due to their assumption that vocabulary tends to converge when various aspects of a product are discussed. Our approach successfully extended upon the work by Titov and McDonald (2008), and Lin and He (2009) who used topic modelling to discover aspects. We managed to separate aspect words from opinion words through the NOA-LDA, whereas previous methods included both opinion words and aspect words in their approach, resulting in the topic model not being very accurate. The NOA-LDA system was also significantly more computationally efficient. In terms of the scalability, run time for the NOA-LDA system was significantly faster than with a raw LDA. This was also demonstrated in both examples, the Hotel reviews performed approximately twice as fast with the NOA-LDA system, and the Headphone reviews performed approximately 15 times faster. The NOA-LDA system was also capable of performing across various domains. We showed this by selecting two examples from different domains.

6.1.2 PLSS

In this dissertation, we also presented an intuitive system, PLSS for Sentiment Analysis that successfully assigned ratings to aspects through a Lexicon-based approach. Our approach employed a customer review orientated Lexicon built to resolve the flaws of previous methods which did not consider how customers used opinionated text to assign ratings. For instance, in the AFINN lexicon, the reviewer needs to use words such as “breathtaking” and “outstanding” in order to allocate 5 stars. This sort of vocabulary is common amongst professional critics, but not amongst the majority of customer reviews. We found that extracting adjectives through POS tagging from customer review titles could allow us to assign a fair and realistic polarity rating to opinionated words. Our intuition behind this is the assumption that adjectives used in review titles typically convey the overall customer sentiment towards a product or service, allowing us to retrieve a sentiment score for each adjective. However, the limitation to our PLSS is that we could not quantifiably measure the accuracy of our method. This potentially weakens our approach until further testing is done.

6.2 Future Work

This work has opened many opportunities for future research. This research could be extended by incorporating more POS into our Aspect Detection approach, such as Noun phrases and Verb phrases. It would also be interesting to develop a method of accurately testing our Sentiment Analysis approach. Finally, we imagine that producing a front-end application would reap the rewards of our approach, thus we aim to continue in that direction.

REFERENCES

- Asuncion, A., Welling, M., Smyth, P. and Teh, Y. (2009). On smoothing and inference for topic models. *UAI '09 Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp.Pages 27-34.
- Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), p.77.
- Blei, D. and Lafferty, J. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1), pp.17-35.
- Blei, D. and Lafferty, J. (2009). A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1), pp.17-35.
- Blei, D., NG, A. and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, pp.993-1022.
- BrightLocal. (2017). *Local Consumer Review Survey | Online Reviews Statistics & Trends*. [online] Available at: https://www.brightlocal.com/learn/local-consumer-review-survey/#Q6.
- Chang, J., Gerrish, S., Boyd-Graber, J., Wang, C. & Blei, D. (2009), Reading Tea Leaves: How Humans Interpret Topic Models, *in* 'Neural Information Processing Systems (NIPS)'.
- Chang, Jonathan & M. Blei, David. (2009). Relational Topic Models for Document Networks. Journal of Machine Learning Research - Proceedings Track. 5. 81-88.
- Chen, K. (2017). *Latent Semantic Analysis & Topic Models.*
- Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, 90(432), p.1313.
- Dave, K., Lawrence, S. and Pennock, D. (2003). Mining the peanut gallery. Proceedings of the twelfth international conference on World Wide Web - WWW '03.
- Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3), pp.411-436.
- Fan and Fuel (2017). *No online customer reviews means BIG problems in 2017 | Fan and Fuel*. [online] Fan and Fuel. Available at: https://fanandfuel.com/no-online-customer-reviews-means-big-problems-2017.
- Feinerer, I. and Hornik, K. (2018). tm: Text Mining Package. R package version 0.7-4. https://CRAN.R-project.org/package=tm.

- Finn Årup Nielsen. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. CoRR, abs/1103.2903. URL http://arxiv.org/abs/ 1103.2903.
- Ganesan, K. and Zhai, C. (2011). Opinion-based entity ranking. *Information Retrieval*, 15(2), pp.116-150.
- Gilks, W., Richardson, S. and Spiegelhalter, D. (1998). *Markov chain Monte Carlo in practice*. Boca Raton: Chapman & Hall/CRC.
- Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Supplement 1), pp.5228-5235.
- Grün, B. and Hornik, K. (2011). topicmodels: AnRPackage for Fitting Topic Models. Journal of Statistical Software, 40(13).
- Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics -.
- Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*.
- Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04.
- Jakob, N. and Gurevych, I. (2010). Extracting opinion targets in a single- and cross-domain setting with conditional random fields. *EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp.Pages 1035-1045.
- Jan Wijffels (2018). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. R package version 0.6. https://CRAN.R-project.org/package=udpipe.
- Kim, S. and Hovy, E. (2004). Determining the sentiment of opinions. Proceedings of the 20th international conference on Computational Linguistics - COLING '04.
- Klinger, R. and Friedrich, C. (2018). User's Choice of Precision and Recall in Named Entity Recognition. *International Conference RANLP 2009*, pp.192–196.
- Lafferty, John & Mccallum, Andrew & Pereira, Fernando. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282-289.

- Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), pp.211-240.
- Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 530–539.
- Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*.
- Luca, Michael. (2011). Reviews, Reputation, and Revenue: The Case of Yelp.Com. SSRN Electronic Journal.
- Lui, B. and Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. *Mining Text data*, pp.415-463.
- McAuley, J. and Leskovec, J. (2018). *Amazon review data*. [online] Jmcauley.ucsd.edu. Available at: http://jmcauley.ucsd.edu/data/amazon/.
- Melville, P., Gryc, W. and Lawrence, R. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Mimno, D.M., Wallach, H.M., Talley, E.M., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. *EMNLP*.
- Minka, Thomas & Lafferty, John. (2002). Expectation-Propagation for the Generative Aspect Model. Uncertainty in Artificial Intelligence (UAI).
- Mittelhammer,R.C., Judge,G.G. and D.J.Miller. (2000). Econometric Foundations. Cambridge.
- Oneata, D. (2011). Probabilistic Latent Semantic Analysis.
- Paltoglou, G. and Thelwall, M. (2010). A study of information retrieval weighting schemes for sentiment analysis. In Proceedings of Annual Meeting of the Association for Computational Linguistics.
- Pang, Bo & Lee, Lillian & Vaithyanathan, Shivakumar. (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques.
- Pang, Bo & Lee, Lillian. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. 2. 1-135.

- Podium. (2017). [online] Available at: http://learn.podium.com/rs/841-BRM-380/images/Podium-2017-State-of-Online-Reviews.pdf?mkt_tok=eyJpIjoiTUdRM04ySTBOR1ZqTURNNSIsInQiOiJVTktEOXNtTXlpZGFhM29YQUFyNXJZWXpNRGhLTUpYVk5nSWdcL0RPMmcwcWdjaFRlazRiMlU5ZDlcL01DMVJBNVdLVHNLYUs0eEM5Uko1dkRCdVZoRHFVbzNDMjNUdTlBT1pCT2t6cHpma3ZTUWNNSnlab3RNblhUTW5Uem5PdWk0In0%3D.
- Popescu, A. and Etzioni, O. (2005). Extracting Product Features and Opinions from Reviews. *Natural Language Processing and Text Mining*, pp.9-28.
- Powell, D., Yu, J., DeWolf, M. and Holyoak, K. (2017). The Love of Large Numbers: A Popularity Bias in Consumer Choice. *Psychological Science*, 28(10), pp.1432-1442.
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Raftery, A., Newton, M., Satagopan, J. and Krivitsky, P. (2007). Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity. BAYESIAN STATISTICS, 8, pp. 1–45.
- Rinker, T. and Kurkiewicz, D. (2017). pacman: Package Management for R. version 0.4.6. *University at Buffalo. Buffalo, New York*. [online] Available at: http://github.com/trinker/pacman [Accessed 20 Sep. 2018].
- Schouten, K. and Frasincar, F. (2016). Survey on Aspect-Level Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3), pp.813-830.
- Shu, Lei & Xu, Hu & Liu, Bing. (2017). Lifelong Learning CRF for Supervised Aspect Extraction.
- Silge, J. and Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Journal of Open Source Software*, 1(3), p.37.
- Spannbauer, A. and White, B. (2017). lexRankr: Extractive Summarization of Text with the LexRank Algorithm. R package version 0.5.0. https://CRAN.R-project.org/package=lexRankr .
- Statista (2018). *Global e-retail growth rate 2021 | Statistic*. [online] Statista. Available at: https://www.statista.com/statistics/288487/forecast-of-global-b2c-e-commerce-growt/.
- Statista. (2018). Global e-commerce share of retail sales. [online] Available at: https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/ .

- Stephens-Davidowitz, S. (2017). *Everybody lies.*
- Tan, S., Wang, Y. and Cheng, X. (2008). Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08.
- Teh, Y., Newman, D., & Welling, M. (2006). A collapsed variational bayesian inference algorithm for latent dirichlet allocation.
- Titov and R. McDonald (2008). Modeling online reviews with multi-grain topic models. In Proceedings of the 17h International Conference on World Wide Web.
- Wei, C., Chen, Y., Yang, C. and Yang, C. (2009). Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and e-Business Management*, 8(2), pp.149-167.
- Wickham, H. (2016). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2. [online] Available at: https://CRAN.R-project.org/package=rvest [Accessed 20 Sep. 2018].
- Wickham, H., Francois, R., Henry, L. and Müller, K. (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr.