UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Aspect-based sentiment analysis using topic modelling on student evaluations

by

Jana Du Toit

Submitted in partial fulfilment of the requirements for the degree

MSc in Advanced Data Analytics

In the Department of Statistics

In the Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

29 November 2024

I, *Jana Du Toit*, declare that this mini-dissertation, which I hereby submit for the degree Masters in Advanced Data Analytics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:

Date: 29 November 2024

ABSTRACT

Aspect-Based Sentiment Analysis (ABSA) is a Natural Language Processing (NLP) task that focuses on identifying and extracting sentiment related to specific aspects or components of various subjects, including but not limited to products or services. In ABSA, the process typically involves several steps. First, aspects or features relevant to the product or service are identified from the text. These aspects could encompass specific attributes, functionalities, or components. Next, sentiment analysis is performed to determine the polarity (positive, negative, or neutral) associated with each aspect based on the context within the sentence or document. Finally, the results are aggregated to provide an overall sentiment for each aspect.

This mini-dissertation investigates a proposed novel approach for aspect-based sentiment analysis using topic modelling on student evaluation data from the Department of Statistics provided by the University of Pretoria. Using ABSA in the higher educational field is significant since it provides insights on how students view certain aspects. These insights are useful for lecturers, as well as the Head of the Department or even the Dean, because they can make certain decisions based on the insights.

The mini-dissertation utilises topic models for aspect extraction. Among these, the Latent Dirichlet Allocation (LDA) topic model is widely recognised. However, literature indicates that the LDA model performs better on longer texts, such as newspaper articles or e-books, rather than shorter texts like tweets. Since the student evaluations used in this research are short texts, the LDA model may not be the most suitable. Therefore, two alternative topic models, the Biterm Topic Model (BTM) and the Dirichlet Multinomial Mixture model (DMM), which are designed for short texts, are also applied to the data. These three topic models are applied in conjunction with an automatic text summarisation method for aspect extraction. As expected, the LDA topic model did not perform as well as the BTM an DMM models. Analysing the results from the BTM

and DMM models, it was evident that the coherence scores from the BTM model were higher than the DMM, which indicates that the BTM model has a better ability to capture the underlying topics and relationships within the data compared to the DMM. After the topic modelling was applied, two sentiment analysis methods, the Multinomial Naïve Bayes method, which is a machine learning technique, and the VADER method, which is a lexicon-based approach, were applied to the educational data. When these two methods were applied to the data it was found that the Multinomial Naïve Bayes approach produced sentiments that were skewed to the negative side. On the other hand, the VADER method produced sentiments that were more evenly spread between positive, neutral and negative sentiments. Therefore, the VADER method was the preferred method. These findings underscore the importance of selecting an appropriate topic modelling approach and sentiment method for aspect-based sentiment analysis tasks.

Key insights and recommendations from analysing the student evaluation data using the proposed new approach to aspect-based sentiment analysis highlight several improvements that the lecturers could consider. These include incorporating pre-recorded videos into the curriculum to accommodate various learning preferences, establishing a peer-review system to reduce errors in assignments and tests, and decreasing the number of pre-class and post-class tests for senior students to better manage their workload. Additionally, customising support and resources to address the specific needs of different student groups and enhancing communication channels between students and faculty to ensure student queries are effectively addressed are also recommended. These recommendations aim to improve the overall learning experience and meet the diverse needs of students.

ACKNOWLEDGEMENTS

I would like to sincerely thank my supervisor, Dr. J. Mazarura. I could not have asked for a better supervisor to walk this journey with me since my honours year. Even though she is no longer working at the university, she continued to support and encourage me.

I would also like to thank my co-supervisor, Dr. I. Fabris-Rotelli. She stepped in for Dr. J. Mazarura when she left the university and has been another source of encouragement and support. I really appreciate the time and effort that both women have put into this research.

I would like to thank my friends, family, and colleagues for their support throughout this journey. Alissa and Carine, we started our academic journeys together, and since day one, you have been there for me through my ups and downs. I am truly grateful for the support you have shown me over the years. Marielle, thank you for always backing my choices and supporting me, even from Cape Town. You are truly the best sister anyone could ask for.

I would like to express my deepest appreciation to my parents. Thank you for supporting me financially, encouraging me, and believing that I can achieve anything I set my mind to. I also want to thank my husband for supporting me throughout this journey. Thank you for listening every time I explained my research, for encouraging me when I did not want to work, and for believing in my dreams.

Lastly, I want to give all my thanks to the Heavenly Father. He has placed people in my life, knowing that I cannot do this alone. All the honour goes to Him, because without Him, nothing would be possible. He has guided and blessed me every step of the way.

The student feedback data that is used in this mini-dissertation has been approved under ethics code: NAS124/2023.

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

LIST OF ABBREVIATIONS

ABSA	Aspect-Based Sentiment Analysis
ADM-LDA	Aspect Detection Model with LDA
AKL	Automated Knowledge LDA
ASUM	Aspect Sentiment Unification Model
BERT	Bidirectional Encoder Representations from Transformers
BERTopic	Bidirectional Encoder Representations from Transformers Topic Model
BTM	Biterm Topic Model
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DMM	Dirichlet Multinomial Mixture
ELDA	Enriched LDA
GK-LDA	General Knowledge LDA
GPM	Gamma Poisson Mixture
GSDMM	Gibbs Sampling DMM
JABST	Joint Aspect-Based Sentiment Topic
JAST	Joint Aspect-Based Sentiment Topic
JST	Joint Sentiment Topic
LAST	Lifelong Aspect-Based Joint Sentiment Topic
LDA	Latent Dirichlet Allocation
LocLDA	Local LDA
LSTM	Long Short-Term Memory
MaxEnt	Maximum Entropy
MaxEnt-LDA	Maximum Entropy LDA

MC-LDA	LDA With M-Set and C-Set
MOOC	Massive Open Online Courses
NLP	Natural Language Processing
PDMM	Poisson DMM
PLSA	Probabilistic Latent Semantic Analysis
SA-ASM	Seller-Aided Aspect-Based Sentiment Model
SA-PSM	Seller-Aided Product-Based Sentiment Model
SS-LDA	Sentence Segment LDA
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency
VADER	Valence Aware Dictionary and sEntiment Reasoner
W2VLDA	Word2Vec LDA

CHAPTER ONE

INTRODUCTION

Aspect-based sentiment analysis (ABSA) is an advanced technique derived from sentiment analysis, a key component of Natural Language Processing (NLP) (Hajrizi and Nuçi, 2020). Sentiment analysis involves computationally identifying and categorising opinions expressed in text to determine whether the sentiment towards a particular topic, product, etc., is positive, negative, or neutral (Bhandari and Ghosh, 2016). It serves as a valuable tool for businesses to gauge customer sentiment regarding their brands and products through platforms like social media or surveys. Sentiment analysis has three different levels of granularity: document level, sentence level and aspect level. At the document level, the focus is to find the overall sentiment of the document. The sentence level determines what the sentiment of a sentence is, and lastly, the aspect level is used to determine what the sentiment around a specific aspect is (Behdenna et al., 2016). Therefore, ABSA goes beyond document or sentence level analysis to identify specific aspects or features of a product or service being discussed in text and their corresponding sentiments. For instance, in a hotel review, ABSA can pinpoint aspects like *location, cleanliness, service,* and *price*, assigning sentiments to each aspect based on customer feedback. While ABSA provides detailed insights for both businesses and consumers, its benefits extend beyond just business applications.

The higher education field could also benefit from these insights (Sivakumar and Reddy, 2017). Students often express their feelings about specific aspects through questionnaires, which sometimes include a free text section for additional comments. Using ABSA, these comments can offer supplementary information beyond the questionnaire data, revealing students' sentiments about specific aspects. These valuable insights can assist lecturers in adjusting their approaches, whether by making changes or maintaining effective practices (Hajrizi and Nuçi, 2020).

1.1 MOTIVATION

Whilst many different methods have been proposed for ABSA, not all of them are applicable to our research. Some researchers use convolutional neural networks for ABSA (Schmitt et al., 2018). Schmitt et al. (2018) used labelled customer reviews to train the neural network to identify aspects and their corresponding sentiments. Since the student evaluation data that is used in this mini-dissertation is not labelled, the convolutional neural network will not be a viable option. Another method that has been used is long short term memory networks (Al-Smadi et al., 2019), which is a deep learning neural network. This is a supervised learning method that requires labelled data to train the network. As with the case of the convolutional neural networks, long short term memory networks will not work on the student evaluation data since the data is not labelled. Other researchers use generative pre-trained transformers (GPT) (Chumakov et al., 2023) to perform ABSA. However, some companies, like OpenAI, make their GPT models available as open-source (Chen et al., 2021). The student evaluation data is private, so using open-source models for ABSA would violate data confidentiality.

In this mini-dissertation, we aim to apply an ABSA model to student evaluations within the Department of Statistics at the University of Pretoria. Given the lack of labelled data and the confidential nature of the dataset, neural network methods and GPT-based approaches are unsuitable.

A promising alternative is topic modelling. Anoop and Asharaf (2018) used the Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003) to extract topics from their data and manually mapped them to specific aspects. They then calculated the polarity score of each aspect-specific sentence, determining sentiment based on whether the score was positive, negative, or neutral. The accuracy of this method ranged between seventy four and eighty one percent for different aspects. The simplicity of this aspect-based sentiment analysis method and its ability to handle unlabelled data make it appealing for our purposes. However, the manual aspect mapping is time-intensive. This mini-dissertation is inspired by the approach of Anoop and Asharaf but introduces some beneficial changes. Our focus is on short pieces of text from student evaluations, for which the LDA topic model might not be optimal. Therefore, we will apply two short text topic models, the Biterm Topic Model (BTM) (Yan et al., 2013) and the Dirichlet Multinomial Mixture (DMM) model (Yin and Wang, 2014). Additionally, we will employ an automatic text summarisation method to map the aspects, aiming to enhance efficiency and accuracy.

1.2 CONTRIBUTIONS

This mini-dissertation introduces a novel approach to ABSA with a special focus on higher education student evaluation data.

This research proposes using a topic model combined with an automatic text summarisation method for aspect extraction, alongside a sentiment analysis method, to create an aspect-based sentiment analysis framework. The main contributions of the research are summarised as follows:

1. Using short text topic models, BTM and DMM, instead of the LDA model to improve topic extraction on the student evaluation data.

2. Introducing an automatic text summarisation method for aspect extraction and labelling, which solves the problem that Anoop and Asharaf experienced with manually mapping topics to aspects.

3. Analysing student feedback from the University of Pretoria and gaining additional insights through aspect-based sentiment analysis. These insights give the lecturers a clearer understanding of how students feel about specific aspects.

1.3 MINI-DISSERTATION STRUCTURE

The rest of the mini-dissertation is structured as follows:

- **Chapter 2: Literature Review**
  This chapter explores the literature available on aspect-based sentiment analysis. It gives insights into what research has been done in terms of aspect extraction, different methods of sentiment analysis and how this has been applied to the educational field.

- **Chapter 3: Topic Modelling**

  Topic modelling is a key component in the proposed approach. This chapter introduces different topic models that are used with longer or shorter pieces of texts.

- **Chapter 4: Sentiment Analysis**

  Sentiment analysis is another key component to the aspect-based sentiment analysis approach that is used in this mini-dissertation. Sentiment analysis can be divided into two main groups, namely machine learning approaches and lexicon-based approaches. These different approaches are discussed in this chapter.

- **Chapter 5: Proposed Methodology**

  This chapter describes the proposed method and showcases its application on a dataset of restaurant reviews, demonstrating how the various elements work together.

- **Chapter 6: Aspect-Based Sentiment Analysis on Student Review Data**

  The proposed aspect-based sentiment analysis method will be applied to the student evaluations dataset and the results from the analysis are discussed in this chapter.

- **Chapter 7: Conclusion**

  This chapter concludes the mini-dissertation by summarising the findings of all experiments conducted.

CHAPTER TWO

LITERATURE REVIEW

Aspect-based sentiment analysis (ABSA) consists of two main components: aspect extraction and sentiment analysis. Aspect extraction identifies and extracts the specific aspects that people evaluate or discuss when reviewing a product or service. For example, reviews of a newly released phone might focus on aspects such as *price*, *battery*, *camera*, or *size*. Various methods exist for performing aspect extraction, and in this mini-dissertation, we will focus on topic modelling as the chosen approach.

The chapter will first introduce topic models and the relevant literature, followed by a discussion of how topic modeling is applied specifically for aspect extraction. Various sentiment analysis methods will also be explored, with a review of the literature surrounding these techniques. Additionally, we will discuss the body of work on ABSA as a whole, with a particular emphasis on its application within the higher education domain.

2.1 TOPIC MODELS

A topic model is a type of probabilistic generative model and an active field of research in computer science, particularly in text mining and information retrieval (Liu et al., 2016). Probabilistic latent semantic analysis (PLSA), developed by Hofmann (1999), is one of the earliest topic models. In this method, each word in a document is treated as a sample from a mixture model. The components of this mixture, which are multinomial random variables, serve as representations of different topics. Consequently, every word in a document can originate from a blend of topics, rather than being tied to a single topic. Every document is reduced to a probability distribution over a predetermined set of topics, represented as a list of *mixing proportions* for these mixture components. These mixing proportions indicate the degree to which each topic contributes to the document. For example, if a document discusses both education and technology, it might be composed of seventy percent from an "education" topic and thirty percent from a "technology" topic. Thus, rather than being associated with a single topic, a document is expressed as a weighted combination of multiple topics. While Hofmann's work contributes to probabilistic text modelling, it lacks a probabilistic model at the document level (Blei et al., 2003). Specifically, each document is represented as a list of numbers that indicate the mixing proportions for different topics. However, PLSA lacks a complete generative probabilistic model for these numbers. This absence of a generative model leads to two significant issues: parameter growth and overfitting, where the number of parameters grows linearly with the corpus size, causing overfitting (Blei et al., 2003). Another issue is assigning probabilities to new documents (Blei et al., 2003). Blei et al. developed the Latent Dirichlet Allocation (LDA) model to address these issues. LDA is a generative probabilistic model for collections of discrete data such as text corpora. It is a three-level hierarchical Bayesian model, where the levels are divided into corpus, document, and words (Blei et al., 2003; Hoffman et al., 2010). These levels correspond with a hierarchical framework where every item in a collection is represented as a limited combination across a foundational set of subjects. Additionally, every topic is represented as an infinite mixing across a set of topic probabilities. LDA addresses the issues of PLSA by introducing a hierarchical Bayesian model with prior distributions. Unlike PLSA, where the number of parameters grows linearly with the corpus size, LDA incorporates priors that prevent overfitting. LDA's generative process allows it to handle new documents by estimating its topic proportions based on the learned topic distributions. The LDA model has been used in many applications, however, researchers have identified sparsity issues, which is when the co-occurrence information from short texts is extremely limited when applying LDA to shorter pieces of text (Albalawi et al., 2020; Wu et al., 2020). The limited word count in short texts makes it challenging for the LDA model to establish meaningful relationships between words (Qiang et al., 2017).

Some researchers aggregated shorter pieces of texts (Weng et al., 2010; Quan et al., 2015;

Jamadi Khiabani et al., 2020) before applying the LDA model to solve the sparsity issue. Other researchers explored topic models specifically designed for short pieces of text such as Yan et al. (2013) who introduced the Biterm Topic Model (BTM). A biterm refers to an unordered pair of words that co-occur within a brief context. In this approach, topics are extracted from short texts by directly modelling the generation of biterms across the entire corpus. The underlying data generation process assumes that the corpus consists of a mixture of topics, and each biterm is drawn from a specific topic. Unlike traditional document-level topic models, the BTM explicitly focuses on modelling word co-occurrence patterns (i.e. biterms) to enhance topic learning. By utilising aggregated patterns from the entire corpus, the BTM addresses the challenge of sparse patterns at the document level through leveraging information from the whole corpus.

Another topic model for short text is the Dirichlet Multinomial Mixture model (DMM) proposed by Nigam et al. (2000). A mixture model is a statistical approach that presumes the analysed data originates from a blend of multiple distributions, with each distribution representing a distinct subgroup or component within the entire population (Baxter, 2010). The DMM serves as a probabilistic generative model for documents. The generative process is predicated on two fundamental assumptions: first, that documents are produced using a mixture model; and second, that mixture components and clusters correspond one to one. The collapsed Gibbs Sampling Dirichlet Mixture Model (GSDMM) (Yin and Wang, 2014) leverages collapsed Gibbs Sampling (Casella and George, 1992; Hollander and Sethuraman, 2001), which simplifies the inference process. Instead of explicitly sampling latent variables (like cluster assignments), GSDMM integrates them out analytically. This integration reduces the number of sampling steps and improves convergence. GSDMM can automatically infer the optimal number of clusters, copes well with sparse short texts, and identifies representative words for each cluster. Another limitation of the DMM model is that it assumes short texts contain only one topic, which can be too restrictive in some cases. To address this limitation, Li et al. (2017) proposed the Poisson Dirichlet Multinomial Mixture Model (PDMM) which allows each short text to be associated with a small number of topics (e.g. 1 to 3 topics). The PDMM relaxes the strict single-topic assumption and provides more flexibility.

Other approaches for short text topic modelling include the Bidirectional Transformers (BERT) model (Devlin et al., 2018), the Gamma Poisson Mixture model (GPM) (Mazarura et al., 2020), the Neural Topic Model (NTM) (Zhao et al., 2021), the BERTopic model (Grootendorst, 2022) and the TextNetTopics model (Voskergian et al., 2023).

2.2 ASPECT EXTRACTION

Aspect extraction is the task of identifying and extracting terms relevant for opinion mining and sentiment analysis, for example terms for product attributes or features. There are different ways that aspects can be extracted such as frequency-based, supervised machine learning or unsupervised machine learning (Bhamare et al., 2019; Yadav et al., 2021). Frequency-based methods rely on calculating the frequencies of nouns and noun phrases to identify important aspects or features in textual data such as term frequency inverse document frequency (TF-IDF) (Sparck Jones, 1972). Supervised machine learning is an approach where a model learns from labelled data to make predictions or classify new instances. In the context of aspect extraction, supervised learning can be used to automatically identify and classify aspects within textual data. Conditional Random Fields (CRF) (Lafferty et al., 2001) is an example of a supervised learning method used for sequence labelling tasks, including aspect extraction. CRF models take into account the sequential nature of text (e.g., words in a sentence) and capture dependencies between adjacent words. Unsupervised machine learning, on the other hand, aims to identify and extract aspects from text data without the use of labelled training data such as topic modelling, where Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a popular topic model.

The introduction of the Joint Sentiment/Topic (JST) model by Lin and He (2009) extended the capabilities of the LDA model by simultaneously detecting topics and their associated sentiments. Unlike the LDA model, which consists of three hierarchical layers, the JST model incorporates an additional fourth layer. In particular, the JST model places a sentiment layer in between the topic and document layers. As a result, the JST model is essentially a four-layer framework in which documents are linked to sentiment labels, topics to sentiment labels, and words to both topics and sentiment labels. One advantage of the JST model over some sentiment analysis models is that it operates as a completely unsupervised model, which can be advantageous in cases where data cannot be easily labelled or trained (Lin and He, 2009). However, this method also has limitations. Notably, the JST model represents documents as a bag-of-words, disregarding the order of words (Lin and He, 2009). The order of the words is important because the model can, for example, give a positive sentiment to the phrase *'not good food'*, which actually needs to be given a negative sentiment.

Another model, called Local LDA (LocLDA), was proposed as an unsupervised approach for aspect extraction from review texts (Brody and Elhadad, 2010). LocLDA focuses on localising the data by treating each phrase as an independent document. It leverages the widely used LDA implementation, GibbsLDA++ (Phan and Nguyen, 2007), to extract aspects from the localised data. Comparisons between aspects identified by LocLDA and manual identification show that LocLDA achieves higher accuracy (Brody and Elhadad, 2010).

In 2010, Zhao et al. proposed the MaxEnt-LDA hybrid model which aims to automatically separate aspect and opinion words in customer reviews. This model combines LocLDA with a supervised Maximum Entropy model to identify aspects and opinions within the reviews. By integrating the Maximum Entropy model, the MaxEnt-LDA model effectively leverages syntactic features to aid in the separation of aspect-related words and opinion-related words. Comparative analysis between the MaxEnt-LDA and LocLDA models reveals that, while both models identify similar aspects, the MaxEnt-LDA can separate aspect words and opinion words (Zhao et al., 2010).

The Aspect Detection Model (ADM-LDA), based on LDA, was presented by Bagheri et al. (2014) as another aspect extraction technique. ADM-LDA is an unsupervised topic modelling technique designed for identifying aspects from review sentences in a sentiment analysis system. It enhances LDA by considering the review sentence structure and extracting corpus-based characteristics through the utilisation of document metadata and wording. While similar to the LDA model in that it uses a hierarchical generative model to connect text parameters, ADM-LDA does not treat documents as a bag-of-words (Wallach, 2006). A bag-of-words is a technique that treats each document as a collection of words, disregarding the order and context in which they appear. Comparative analysis with LDA demonstrates that ADM-LDA achieves better aspect extraction (Bagheri et al., 2014).

Topic models optimise specific objectives during training. However, this objective does not directly consider human interpretability or coherence. When assessing topic quality, we rely on human intuition. Coherent topics should make sense to people, be interpretable, and capture meaningful themes. Unfortunately, fully unsupervised models may prioritise statistical patterns over semantic coherence. Incoherent topics arise when the model assigns words randomly or based solely on co-occurrence statistics. As a result, topics lack clear themes or meaningful connections (Chang et al., 2009). To address this issue, Chen et al. (2014) proposed the Automated Knowledge LDA (AKL) model. This method mines prior knowledge from a large amount of web data and utilises topic models to generate more coherent aspects. AKL consists of two steps: first, a technique is introduced to automatically learn knowledge by applying LDA to the data, clustering the topics, and mining frequent patterns; second, a new topic model with an advanced inference mechanism incorporates the learned knowledge in a manner capable of handling possibly incorrectly learned knowledge.

The enriched LDA (ELDA) has been proposed as an alternative model to minimise incorrectly learned knowledge (Shams and Baraani-Dastjerdi, 2017). ELDA enhances aspect precision by integrating word co-occurrence information into the standard Latent Dirichlet Allocation (LDA) model. By analysing which words tend to co-occur, ELDA captures semantic relationships. This process automatically generates knowledge, preventing incorrect or overly general aspect extraction. ELDA ensures that the extracted aspects are meaningful and relevant to the topic.

The Turkish language, for example, presents a challenge for existing aspect-based sentiment models, as a sentence in English may be expressed as a single word in Turkish (Ozyurt and Akcayol, 2021). To address this issue, the Sentence Segment LDA (SS-LDA) model was proposed (Ozyurt and Akcayol, 2021). This model first identifies the polarity of sentences using a sentiment dictionary and then segments the sentences into distinct parts. Segments pertaining to the same product aspects across different sentences are grouped together. By assigning topics to sentence segments instead of single words, SS-LDA improves the accuracy of grouping. Comparative evaluations with LDA, Sentence LDA, and the Biterm Topic Model indicate that SS-LDA outperforms them in aspect extraction, which is crucial for sentiment analysis (Ozyurt and Akcayol, 2021).

Another study explored the use of LDA as a feature extraction method for Indonesian hotel review data (Hidayati, 2023). The research proposed integrating LDA with two word embeddings-Word2Vec and Doc2Vec (Mikolov et al., 2013). Word2Vec is a neural network-based model that learns vector representations of words by predicting word context, while Doc2Vec extends this approach to represent entire documents, capturing semantic similarities between texts. Through various comparisons, it was found that LDA performed better when combined with these embeddings. Notably, the integration with Doc2Vec yielded more accurate feature extractions than with Word2Vec.

Within the literature, it is common for researchers to use the LDA topic model in different ways to carry out aspect extraction. It shows that the use of the LDA topic model is a very popular approach. However, since the student evaluation data used in this mini-dissertation are short pieces of text, we propose that a short text topic model could be a better fit than the LDA model.

2.3 SENTIMENT ANALYSIS

Sentiment analysis, or opinion mining, assigns an opinion or emotional label to text (Kausar et al., 2019). Most often this label indicates polarity, whether the text expresses a positive or negative opinion (Stine, 2019). Sentiment analysis can be done using different methods, such as lexicon-based approaches or machine learning approaches (Wankhade et al., 2022). Lexicons are collections of tokens, each assigned a predetermined score indicating the neutral, positive, or negative nature of the text (Kiritchenko et al., 2014). Scores are based on polarity, typically ranging from 1 for highly positive to -1 for highly negative. A machine learning approach for sentiment analysis involves training a model using machine learning algorithms based on training data.

2.3.1 MACHINE LEARNING

Machine learning techniques, such as Naïve Bayes (Kibriya et al., 2005), Support Vector Machines (Steinwart and Christmann, 2008), and Maximum Entropy (Jaynes, 1957), are commonly used for categorising the polarity of sentiment. Popular techniques for ABSA also include Random Forests (Al Amrani et al., 2018).

NAÏVE BAYES

Naïve Bayes is a popular and straightforward algorithm used for classification tasks (Medhat et al., 2014; Birjali et al., 2021). It is based on Bayes' Theorem, which is a mathematical expression for calculating the probability of an event based on prior knowledge of conditions related to the event. A key component of the Naïve Bayes algorithm is the bag-of-words feature extraction technique (Jurafsky and Martin, 2014). Thus, the presence of a word in a document is considered independently of the presence of other words. Despite this simplifying assumption, Naïve Bayes performs remarkably well in various real-world applications due to its robustness and efficiency (Raschka, 2014). This effectiveness is attributed to the algorithm's ability to handle large datasets and its simplicity in implementation, making it a preferred choice in many practical scenarios (Raschka, 2014).

Troussas et al. (2013) conducted an experiment to classify the polarity of Facebook statuses using different classification techniques. The accuracies of the different classifiers were compared and it was found that the Naïve Bayes had the best performance, based on the accuracy of the model.

MAXIMUM ENTROPY

Maximum entropy (MaxEnt) is a classification method that is also known as a conditional exponential classifier (Berger et al., 1996). It is different from the Naïve Bayes classifier in that it does not make any assumptions about the relationships between features. Instead, it uses encoding to convert labelled feature sets into vectors, which are used to calculate weights for each feature. These weights are combined to determine the most likely label for a feature set, making it a flexible and powerful tool for sentiment analysis.

SUPPORT VECTOR MACHINES

Support Vector Machines (SVMs) are a popular supervised algorithm used for classification (Brereton and Lloyd, 2010). SVMs have proven suitable for text classification, and since their introduction, they have become a common sentiment classification algorithm, used to categorise text into positive, negative, or neutral sentiment (Joachims, 2005). The main goal of the SVM classifier is to find the optimal hyperplane that separates the classes with the largest margin, producing fewer errors and leading to better generalisation performance on new data (Cortes and Vapnik, 1995). Rana and Singh (2016) analysed the sentiment of movie reviews using two different classification methods. They applied SVM and Naïve Bayes on the data. After comparing the results from the two methods they found that the SVM produced more accurate results than Naïve Bayes.

RANDOM FOREST

Random Forests, an ensemble approach based on decision trees, were created by Breiman (2001). This supervised classification algorithm constructs multiple decision trees using random subsets of the training data and features. Each decision tree is trained independently and makes predictions on new instances. The final prediction of the Random Forest classifier is determined by aggregating the predictions of all the individual trees, typically by taking the mode among the predictions (Attanasi and Coburn, 2020). Fauzi (2018) applied Random Forests for sentiment classification in the Indonesian language and found that it had a good classification performance.

Each of the approaches offer unique advantages. Maximum Entropy does not assume independence between features, allowing it to model more complex relationships, which can be beneficial for nuanced sentiment analysis. However, it requires more computational resources. Support Vector Machines excel in high-dimensional spaces (Guyon et al., 2004) and are effective for text classification, aiming to find the optimal hyperplane that maximises the margin between sentiment classes, thus improving generalisation. They can be computationally intensive but are powerful for distinguishing subtle differences in sentiment (Pang et al., 2002). Random Forests are robust and handle overfitting well by averaging multiple decision trees, making them suitable for datasets with a mix of features. They are also relatively easy to tune and interpret, providing good performance in sentiment classification tasks (Sanchez-Medina, 2024). Naïve Bayes, however, is often favoured for its simplicity and efficiency, particularly with large text datasets. Its assumption of feature independence, while sometimes a limitation, often works well in practice for text classification tasks (Wickramasinghe and Kalutarage, 2021), making it a strong candidate for sentiment analysis.

2.3.2 LEXICON-BASED APPROACHES

Lexicon-based approaches, also known as knowledge-based approaches, are sentiment analysis techniques that use an opinion lexicon (Turney, 2002). An opinion lexicon is a predefined list of words which associate words with their semantic orientation indicated as positive or negative using scores (Turney, 2002). These scores can be presented in different ways, such as -1, 0, and 1 for negative, neutral, and positive words, respectively, or as a continuous value that represents the sentiment strength. Typically, these scores are found through methods such as manual annotation by experts, statistical analysis of large text corpora, or machine learning algorithms that learn from labelled data (Han et al., 2018). For example, consider the following sentence: *The movie was good, but the ending was horrible.* There are 2 sentiment words *good* and *horrible*. According to a lexicon dictionary, *good* may have a sentiment score of 1.9 and *horrible* may have a score of -2.5, according to a lexicon dictionary. To get the overall sentiment of the sentence the scores of the emotion words are added. Therefore the overall sentiment of the sentence is -0.6, which is smaller than 0 and so the sentiment is negative.

There are two main lexicon-based approach techniques: dictionary-based approaches, and corpus-based approaches (Gupta and Agrawal, 2020). Dictionary-based approaches use pre-existing dictionaries that contain sentiment scores for words while corpus-based approaches involves using statistical methods to derive sentiment scores from a large corpus of text.

In 2005, Liu et al. introduced the Liu lexicon for sentiment analysis, utilising a dictionary of sensitive terms - words or phrases that carry emotional connotations or express opinions. These terms are divided into positive and negative sets. Words are assigned to these sets for sentiment classification. Experiments showed accurate results, but the lexicon's limited word coverage hinders its effectiveness on large datasets, potentially impacting accuracy (Zhang and Liu, 2017).

SentiWordNet (Sebastiani and Esuli, 2006) takes a different approach. It is a lexicon-based method that assigns sentiment scores to words based on their semantic relationships. Over time, SentiWordNet has undergone improvements, with the latest version being SentiWordNet 3.0 (Baccianella et al., 2010). Notably, SentiWordNet 3.0 was compared to its earlier version, SentiWordNet 1.0, and the results indicated improved accuracy. Interestingly, Vu and Le (2017) combined the Liu lexicon with SentiWordNet 3.0 for their research. After conducting experiments, they found that this combined approach yielded more accurate results in terms of sentiment analysis compared to using the individual approaches separately.

Another popular lexicon approach is the Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto and Gilbert, 2014). VADER is a lexicon and rule-based sentiment analysis Python[1] tool that is open source under the MIT license. The VADER lexicon performs exceptionally well in the social media domain, and it is easy to inspect, understand, apply quickly, and extend for various applications. According to Bonta et al. (2019) VADER's performance is of gold-standard quality and has been validated by users. In addition to VADER, TextBlob (Loria et al., 2018) is another lexicon approach used in Python. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, and sentiment analysis.

2.4 ASPECT-BASED SENTIMENT ANALYSIS USING TOPIC MODELS

The previous sections of this chapter reviewed the literature on the individual components of ABSA. This section will examine the literature on ABSA as a whole, with the addition of topic models. In 2011, the Aspect Sentiment Unification Model (ASUM) (Jo and Oh, 2011) was pro-

[1] https://www.python.org/

posed. This model aims to identify the sentiment polarity of different aspects. It uses sentence-LDA (Jo and Oh, 2011), which is an extension of LDA that assumes all words in a sentence are generated from one aspect. The sentence-LDA is then extended to form ASUM, which incorporates aspects and sentiments. However, it was found that the assumption that a sentence only has one aspect could be improved.

In 2016, a gap in existing models for aspect-based sentiment analysis (ABSA) was identified, where models struggled to apply all four subtasks of ABSA (Wang et al., 2016). These subtasks include aspect extraction, which involves identifying the specific aspects or features of a product or service mentioned in the text; opinion identification, which detects the opinions or sentiments expressed about the identified aspects; polarity classification, which classifies the detected opinions as positive, negative, or neutral; and distinguishing between general and aspect-specific opinions, which differentiates between general sentiments about the overall subject and specific sentiments about particular aspects (Wang et al., 2016). This gap prompted the proposal of the Joint Aspect-based Sentiment Topic (JAST) model (Wang et al., 2016). JAST aimed to simultaneously model aspect extraction, opinion identification, polarity classification, and separation of general and aspect-specific opinions through joint aspect and sentiment modelling. The model utilised LDA for aspect modelling and determined the sentiment polarity of each aspect. However, it occasionally misclassified aspect terms as general opinion words and vice versa. To address this limitation, an enhanced version of JAST called the Lifelong Aspect-based Joint Topic (LAST) model was introduced (Wang et al., 2016). LAST incorporates lifelong machine learning, an advanced paradigm in machine learning that continuously evolves by learning from past experiences, accumulating knowledge, and adapting this knowledge to improve future learning and problem-solving capabilities. This allows the model to extract prior knowledge from multiple domains and enhance its performance.

In a separate study the aspect term extraction limitations of ASUM were addressed by incorporating product descriptions in product reviews (Amplayo et al., 2018). Two ASUM-extended models were proposed: the Seller-aided Aspect based Sentiment model (SA-ASM) and the Seller-aided Product based Sentiment model (SA-PSM). SA-ASM placed the topic distribution within the reviews, while SA-PSM placed it within the descriptions. Comparative analysis with JAST and ASUM revealed that SA-ASM achieved the highest performance in sentiment classification and aspect assignment, while SA-PSM demonstrated superior aspect term extraction capabilities.

In another study about products, a method was proposed to assist customers in choosing between similar products (Guo et al., 2018). They combined rule-based aspect extraction with LDA for topic discovery. Sentimental scores were calculated using a lexicon-based approach, and overall product scores were obtained by combining these scores. An improved PageRank algorithm (Page et al., 1999) ranked the products, providing customers with recommendations. The predicted rankings of the method were strongly correlated with the actual sales rankings.

In their study, Ali et al. (2019) explored ways to enhance traffic control and transportation services using sentiment analysis of tweets related to traffic and transportation. Initially, they considered LDA for topic modelling. However, the LDA had several limitations: the topics that were generated included irrelevant features when other transportation-related text was present. Secondly, it produced noisy topics from the short texts and it missed valuable topics because of the limited dataset. Lastly, the LDA did not consider the relation between topics and documents with low-probability words. To address these issues, the authors proposed an ontology and LDA-based topic modelling system. They combined LDA with ontology-based semantic knowledge to extract more accurate topics. Additionally, word embedding techniques transformed words into low-dimensional vectors, and a deep learning classifier determined word polarity.

Wahid et al. (2021) proposed an enhanced approach for sentiment classification. They applied the LDA topic model to Twitter data, resulting in different topics. These topics were then converted into different aspects. Next, they used the linguistic inquiry and word count lexicon to extract sentiments from the new dataset that included these aspects. Finally, SVM, Random Forest, and Naïve Bayes classifiers were employed to classify the sentiments of the aspects into three categories: negative, neutral, and positive. In their work, Pathik and Shukla (2022) introduced an approach for ABSA of textual reviews that is nearly unsupervised. This method leverages minimal labelled data, reducing the need for extensive manual annotation typically required in supervised methods. They combined LDA with linguistic rules to extract aspects from the text. These extracted aspects were then ranked based on their probability distribution values and clustered into predefined categories using domain knowledge. Finally, sentiment scores were computed using SentiWordNet (Baccianella et al., 2010), resulting in an average accuracy of 85% when tested on manually labelled data.

Tang et al. (2019) introduced the Joint Aspect-Based Sentiment Topic (JABST) model, which simultaneously captures diverse aspects and opinions by modelling aspects, opinions, sentiment polarities, and granularities. Within this unified framework, the model addresses several tasks, including aspect extraction, opinion identification, sentiment polarity classification, and the separation of aspect-related and opinion-related words. While the initial JABST model performed well overall, it faced challenges in distinguishing opinions effectively. To address this, the researchers proposed the MaxEnt-JABST model, which leverages supervised learning based on Maximum Entropy principles to achieve better separation of opinions and aspects. Saidi et al. (2022) employed aspect-based sentiment analysis to identify Twitter profiles of terrorists with high cyber social impact. They used the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) topic model to extract features from tweets. These features were then used to train SVM, Naïve Bayes, and logistic regression classifier models for polarity analysis. After identifying the polarity of different features, they distinguished between extremist and non-extremist Twitter profiles. Ali et al. (2021) investigated Marrakech's negative reputation on TripAdvisor using aspect-based sentiment analysis. They proposed a model combining topic modelling and lexicon-based algorithms. LDA was used to find relevant topics based on tourists' interests and hidden tendencies in reviews. Two lexicon-based algorithms, Textblob[2] and Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto and Gilbert, 2014), were applied to analyse sentiments. The study found that both proposed models outperformed the JST model, with Textblob achieving higher accuracy than VADER. Hoang et al. (2019) explored a weakly supervised Aspect-Based Sentiment Analysis (ABSA) using contextual word representations from the pre-trained language model BERT. Their approach involved fine-tuning BERT with additional text, achieving state-of-the-art results on SemEval 2015 (Pontiki et al., 2015) and SemEval 2016 (Nakov et al., 2016) ABSA tasks. SemEval is a series of international workshops focused on NLP research. Its mission is to advance the state of the art in semantic analysis and to develop high-quality annotated datasets for a variety of increasingly complex problems in natural language semantics. Notably, this study was the first to explore weakly supervised ABSA for aspect classification. The proposed model predicted aspects for both in-domain and out-of-domain scenarios, with a sentiment classifier fine-tuned to identify sentiment context using contextual representations. Overall, their combined model outperformed previous state-of-the-art results.

In 2018, Anoop and Asharaf proposed a method that firstly identified different aspects of

[2]This is a Python package with the following link to the documentation: https://textblob.readthedocs.io/en/dev/

product reviews using topic modelling and then secondly performed sentiment analysis on those aspects. They used LDA to find different topics which were mapped to the corresponding aspects of the reviews. The Naïve Bayes classifier was then used to compute the polarity of the different aspects. Lastly, they found the overall sentiment of each review using the aspect-specific sentiments. Their model was compared to an ontology-based approach as well as a linguistic approach and it outperformed both.

As seen in the aspect extraction section, as well as this section, most researchers use the LDA topic model. Therefore, our approach was inspired by the success reported using the method in Anoop and Asharaf (2018). However, instead of using a manual method to map the aspects, we propose using an automatic text summarisation method in addition to short text topic models.

2.5 ASPECT-BASED SENTIMENT ANALYSIS IN HIGHER EDUCATION

ABSA methods have been applied to customer feedback, product reviews and social media posts. However, ABSA can also be applied in the educational field, assessing how students feel about certain aspects. This section will discuss the literature on aspect-based sentiment analysis applied to higher education.

Sivakumar and Reddy (2017) employed ABSA to identify specific aspects (such as teaching methods or resources) within student feedback. By applying machine learning techniques, including Naïve Bayes and Support Vector Machines, they analysed sentiment for each aspect. The approach provides valuable insights for educators to enhance various aspects of education. Their research analysed student feedback from Twitter, identifying seven aspects: teaching, placement, facilities, sports, organising events, fees, and transport. Among these, sports, organising events, and fees received more negative feedback, while facilities had a balanced mix of positive and negative sentiments. Hajrizi and Nuçi (2020) explored ABSA in education, highlighting its potential to see which aspects of certain subjects students evaluate as positive or negative. They outline the ABSA process, emphasising the significant role that data pre-processing plays - removing any noisy data that could affect the outcome of the modelling. They also highlight that the addition of word embeddings improved the performance of some of the methods. The authors note the increasing popularity of neural networks, particularly deep learning models leveraging word2vec and LDA to generate word embeddings and topic vectors. This integration of semantic information, they argue, enhances the ability of the models to understand student sentiment and identify aspects like course content and teaching methods. Despite limitations like the lack of large annotated datasets, their review emphasises the progress made in utilising ABSA and word embeddings to understand student voices and contribute to educational improvements.

Another study revealed a novel framework that can automatically analyse the opinions of students who write reviews of massive open online courses (MOOCs) which are online courses that are open to anyone (Kastrati et al., 2020). The framework uses ABSA, which identifies the specific aspects of the MOOC about which the students are expressing their feelings about, such as the instructor, the content, the difficulty, the platform, etc. Some of the words that are related to the instructor are *knowledge, subject, good, teacher, sir, excellent, informative*. These words are used in a positive way, which could imply that the students are positively disposed towards the instructors. The framework also determines the polarity of their opinions, whether they are positive, negative, or neutral. The framework does not rely heavily on manually labelled data. Instead, it employs CNN and LSTM models for semi-supervised learning, leveraging existing knowledge or rules to label review aspects. Weak labels, which are labels not manually annotated with high precision but generated through automated processes, or existing knowledge, are used to train the model. These labels might be less accurate or noisy compared to manually labelled data. The model then applies this learning to unlabelled reviews, improving its ability to identify and classify aspects in the data.

Bhowmik et al. (2023) noted that teacher evaluation often relies on subjective measures, highlighting the need for more objective and data-driven approaches. They addressed the need for a high-quality dataset by introducing ABSA-Edu, a unique dataset specifically designed for ABSA in teacher evaluation. Unlike generic ABSA datasets, ABSA-Edu leverages 2 million student reviews from American International University-Bangladesh, meticulously labelled by undergraduates across various aspects of teacher performance, such as teaching style, knowledge, fairness, and overall satisfaction. This domain-specific focus and detailed labelling address limitations of previous datasets, enabling more accurate sentiment analysis and refined teacher feedback. Interestingly, when comparing the ratings assigned to reviews and the actual review content, discrepancies were observed. For negative reviews, common words like *good*, *teacher*, and *understand* occurred frequently. This suggests that students comments did not always align with the assigned ratings.

ABSA research within higher education explores various methodologies and datasets to understand student feedback and enhance educational experiences. Existing studies have made progress by utilising machine learning techniques, word embeddings, and domain-specific datasets to analyse student sentiment across different aspects. However, there remains a gap in investigating short text topic models for aspect extraction, in the higher education domain. This gap presents an opportunity for additional research, which will be addressed in this mini-dissertation.

2.6 CONCLUSION

In this chapter, we explored the literature related to aspect-based sentiment analysis. By applying ABSA to student feedback, educational institutions gain valuable insights as students often provide feedback on different aspects of their experiences, such as course content, teaching quality, campus facilities, and administrative processes. ABSA allows us to extract sentiments related to these specific aspects, enabling targeted improvements. The chapter discussed various approaches to ABSA. For our research, we specifically focused on the topic modelling approach for aspect extraction. We made this choice because our data is unlabelled. Topic modelling, particularly using techniques like LDA, helps identify hidden topics within a collection of documents. As is evident in the literature, most ABSA approaches incorporate LDA into their methods. Previous work by Anoop and Asharaf (2018) demonstrated that a simplistic method, using LDA, had great accuracy results. However, their approach involved manual mapping of aspects. In our research, we propose to replace this manual step with an automatic text summarisation method, improving the time efficiency of the model. We also change the topic model from an LDA model to a short text topic model, specifically the DMM and the BTM. The topic model is changed since the student reviews are shorter pieces of text. Our specific focus on the higher education domain stems from the potential impact on student experiences. By analysing student feedback, we can identify areas for improvement, enhance teaching quality, and optimise administrative processes. In the next chapter, specific topic models will be discussed in more detail.

CHAPTER THREE

TOPIC MODELLING

3.1 INTRODUCTION

Topic modelling, widely used in the field of Natural Language Processing (NLP), aims to uncover underlying thematic structures within a text corpus (Bhandari and Ghosh, 2016). It involves algorithms that analyse word and phrase co-occurrences within documents, independently identifying groups of words that characterise these texts. These groups often represent coherent themes or topics. At its core, topic modelling is a statistical method that identifies patterns within a given corpus, unveiling latent topics inherent in the data. This process provides a structured representation of content, empowering researchers, analysts, and practitioners to identify key themes across a range of documents.

In this chapter, and the rest of the mini-dissertation, the following terminology will be used:

1. Corpus: A large and structured set of documents, typically used for linguistic research. In the context of student evaluation data, it refers to the entire collection of student evaluations.

2. Document: A set of sentences within the corpus. For this study, a document represents a single student evaluation.

3. Sentence: A sequence of words within a document that conveys a complete thought. Each student evaluation contains multiple sentences expressing various opinions.

4. Aspect: A specific feature or component in a sentence. In the context of student evaluations, an aspect might be *time, content* or *lecturer*.

Consider the following example for a better understanding of how a typical topic model works. Table 3.1 consists of 8 documents that have been selected from different websites[1]. These sentences fall under four different categories, namely, *dogs, food, golf* and *fashion*.

Table 3.1: Example dataset illustrating topic modelling. Each cell represents a document, with sentences sourced from different websites.

1	A dog's nose print is unique, much like a person's fingerprint.
2	All dogs dream, but puppies and senior dogs dream more frequently than adult dogs.
3	Meat is one of the most widely consumed foods globally, and its production and consumption are increasing year by year
4	The dark chocolate we know and love is made from the roasted beans of the cacao tree and has a ton of interesting facts-many being health benefits.
5	The modern game of golf can be traced to Scotland during the 15th century.
6	The first balls used in playing golf were made from leather balls stuffed with bird feathers.
7	In the United States, the average person owns seven pairs of blue jeans — one for every day of the week!
8	Anna Wintour's first cover of Vogue was so different to the previous editions that the editors thought she had made a mistake

After cleaning the sentences, the documents in Table 3.1 appear as shown in Table 3.2. The cleaning process involves removing common stopwords such as *a*, *and*, and *the*, converting all words to lowercase, and removing any numbers.

[1]https://www.signos.com/food-comparison/beef-vs-chicken-which-is-better
https://www.eatfirst.com.au/en-au/c/blog/fun-facts-about-food
https://www.akc.org/expert-advice/lifestyle/dog-facts/
https://thefactfile.org/golf-facts/
https://www.thefactshop.com/fashion-facts/100-fashion-facts/

Table 3.2: Pre-processed version of the documents from Table 3.1.

1	dog nose print unique much like person fingerprint
2	dog dream puppy senior dog dream frequently adult dog
3	meat one widely consumed food globally production consumption increasing year year
4	dark chocolate know love made roasted bean cacao tree ton interesting fact many health benefit
5	modern game golf traced scotland century
6	first ball used playing golf made leather ball stuffed bird feather
7	united state average person owns seven pair blue jeans one every day week
8	anna wintours first cover vogue different previous edition editor thought made mistake

Next, a document-word matrix is constructed based on the cleaned texts in Table 3.2, where each row represents a document, each column represents a unique word from the corpus (the vocabulary), and each cell contains the frequency of that word in the respective document. This matrix, shown below, is a visual representation of the bag-of-words model:

$$
\begin{array}{c}
 \\ \text{document 1} \\ \text{document 2} \\ \text{document 3} \\ \text{document 4} \\ \\
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{dog} & \text{nose} & \text{print} & \text{unique} & \end{array} \\
\begin{pmatrix}
1 & 1 & 1 & 1 & \cdots \\
3 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\end{array}
$$

The matrix above shows only the first few words and documents for demonstration purposes. In the full matrix, there would be 76 columns (representing each unique word) and 8 rows (one for each document in the corpus). From the document-word matrix above, it can be seen that the first column is the word *dog* which appears once in the first document and 3 times in the second documents, but zero times for the rest of the documents. Secondly, the word *nose* is once in the first document, but not in the rest of the documents. The rest of the matrix follows the same pattern, counting how many times each word appears in each document.

In a typical topic model, this document-word matrix is used to estimate the probability of each word belonging to specific topics. The model outputs groups of words that collectively describe each topic. A topic is represented as a probability distribution across all words in the corpus. Usually, ten words with the highest probabilities of belonging to a topic are selected to describe each topic.

Table 3.3: Example of three topics uncovered by a topic model and the probability that the words belong to each topic.

Topic 1 (dog)	Probability of word in topic 1	Topic 2 (food)	Probability of word in topic 2	Topic 3 (golf)	Probability of word in topic 3
dog	0.23	chocolate	0.37	golf	0.217
fingerprint	0.146	love	0.22	century	0.143
nose	0.124	bean	0.16	game	0.137
person	0.05	tree	0.145	scotland	0.093
much	0.029	fact	0.09	modern	0.033

Table 3.3 shows the top five words with their probabilities of belonging to topics 1, 2 and 3 using a simple topic model.

Lastly, each of the documents are assigned to a topic. Most topic models include a parameter that determines the proportion of each topic within a document. These proportions are then showcased as a document-topic matrix where the rows indicate the documents, the columns indicate the topics and the cells indicate the proportion of the documents belonging to each topic. Continuing with the toy example from Table 3.1, assuming that each document belongs to a single topic, the document-topic matrix looks as follows:

$$
\begin{array}{r c}
 & \begin{array}{cccc} \text{topic 1 (dog)} & \text{topic 2 (food)} & \text{topic 3 (golf)} & \text{topic 4 (fashion)} \end{array} \\
\begin{array}{r} \text{document 1} \\ \text{document 2} \\ \text{document 3} \\ \text{document 4} \\ \text{document 5} \\ \text{document 6} \\ \text{document 7} \\ \text{document 8} \end{array} &
\begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1
\end{pmatrix}
\end{array}
$$

From this matrix, it is shown that document 1 and 2 are assigned to topic 1, for example. Both these documents are about dogs, so this assignment is expected.

3.1.1 NOTATION

In the sections and chapters to follow the following notation will be used:

- C: A corpus, which is a collection of documents.
- D: The number of documents in the corpus, denoted as D. Each document is indexed by d_k, $k = 1, 2, \ldots, D$.
- T: The number of topics, denoted as T. Each topic is indexed by t_l, $l = 1, 2, \ldots, T$.
- w_{i,d_k}: The i^{th} word in document d_k, where i is the position of the word within that document.
- β: The topic-word distribution, where $\beta = [\beta_1, \beta_2, \ldots, \beta_T]$ and β_l, $l = 1, 2, \ldots, T$ is the vector of word probabilities for topic t_l. Each element in β_l corresponds to the probability of a particular word in the vocabulary being associated with topic t_l.
- θ: The document-topic distribution, where $\theta = [\theta_1, \theta_2, \ldots, \theta_D]$ and θ_k, $k = 1, 2, \ldots, D$ represents the distribution of topics within document d_k, and each element in θ_k corresponds to the probability of that topic appearing in the document.
- z: Denotes the vector of topic assignments for the words in the corpus. Specifically, $z_k = [z_{k,1}, z_{k,2}, \ldots, z_{k,n_k}]$ is the vector of topic assignments for the words in document d_k, where $z_{d,i} \in \{t_1, t_2, \ldots, t_T\}$ indicates the topic assigned to the i^{th} word in that document, and n_k represents the total number of words in document d_k.
- ϕ: The corpus-wide topic distribution, where $\phi = [\phi_1, \phi_2, \ldots, \phi_T]$. This denotes the distribution of topics across the entire corpus, representing the overall prevalence of each topic in the corpus.
- γ_β and α: The Dirichlet prior hyperparameters for β and θ, respectively, used to control the sparsity of the topic-word and document-topic distributions.

- $b = (w_i, w_j)$: Represents a biterm, which is a pair of words (w_i, w_j) that frequently appear together within the same document or within the entire corpus. This notation is particularly relevant in models like the Biterm Topic Model (BTM).
- s: A sentiment class, where $s \in \{\text{positive, neutral, negative}\}$.
- N_s: The number of documents belonging to a certain sentiment class.
- $N_{w_i,s}$: The number of times word w_i appears in sentiment class s.
- $|C|$: The number of unique words in the corpus.
- L: The VADER lexicon dictionary.
- o_i: The sentiment score of word w_i using the VADER lexicon.
- O_{comp}: The compound sentiment score of a sentence. This is a normalised sum of all sentiment scores in a sentence.

Using the notation listed, the Latent Dirichlet Allocation (LDA), Biterm Topic Model (BTM), and Dirichlet Multinomial Mixture (DMM) topic models will be discussed below.

3.2 LATENT DIRICHLET ALLOCATION

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a model designed to reveal latent topics embedded within a corpus of text documents. Unlike supervised learning methods, LDA adopts a data-driven approach by identifying latent topics based on observed word co-occurrence patterns. This flexibility allows LDA to uncover hidden themes that may not be apparent through manual categorisation. According to Blei et al. (2003), LDA assumes the following generative process for each document in a corpus C is represented as:

1. Randomly choose T topic distributions:

   For each topic t_l, $l = 1, 2, \ldots, T$ sample $\beta_l \sim \text{Dirichlet}(\gamma_\beta)$[2] , where β_l represents the

[2] The k-dimensional Dirichlet distribution is defined as follows:

$$p(\theta \mid \gamma_\beta) = \frac{\Gamma\left(\sum_{i=1}^{k} \gamma_{\beta_i}\right)}{\prod_{i=1}^{k} \Gamma(\gamma_{\beta_i})} \theta_1^{\gamma_{\beta_1}-1} \cdots \theta_k^{\gamma_{\beta_k}-1}$$

where $\Gamma(x)$ represents the gamma function, $\gamma_\beta = [\gamma_{\beta_1}, \gamma_{\beta_2}, \ldots, \gamma_{\beta_k}]$ is a k-dimensional vector with each component $\gamma_{\beta_i} > 0$, $\theta_i \geq 0$ and $\sum_{i=1}^{k} \theta_i = 1$.

topic-word distribution for topic t. The Dirichlet distribution serves as a prior over the multinomial distributions governing the topic-word distributions.

2. For each document d_k, $k = 1, 2, \ldots, D$:
   Sample $\theta_k \sim$ Dirichlet(α), where θ_k is a document-topic multinomial distribution. The parameter α controls the sparsity of the document-topic distributions.

3. For each word, w_{i,d_k}, in document d_k:
   Sample $z_k \sim$ multinomial(θ_k). This step assigns a topic to each word based on the document's topic distribution. Then sample $w_k \sim$ multinomial(β_{z_k}), where β_{z_k} is the document-topic distribution for topic assignment z_k.

Given the parameters α and γ_β, the joint distribution of a document-topic vector θ, a set of N topics z and a set of M words w is given by

$$p(\theta, z, w|\alpha, \gamma_\beta) = p(\theta|\alpha) \prod_{n=1}^{N} \prod_{m=1}^{M} \Big[p(z_n|\theta) + p(w_m|z_n, \gamma_\beta)\Big]. \tag{3.1}$$

This distribution combines the probability of the topic distribution $p(\theta|\alpha)$, the topic assignment $p(z_n|\theta)$, and the word given the topic $p(w_m|z_n, \gamma_\beta)$. The marginal distribution of a document is found by integrating over θ and summing over z:

$$p(w|\alpha, \gamma_\beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \prod_{m=1}^{M} \sum_{z_n} p(z_n|\theta) + p(w_m|z_n, \gamma_\beta) \right) d\theta. \tag{3.2}$$

To understand how the posterior distribution in LDA is derived, we start with the joint distribution of the document-topic vector θ, a set of topics z, and a set of words w given the hyperparameters α and γ_β. This joint distribution is expressed as:

$$p(\theta, z, w|\alpha, \gamma_\beta) = p(\theta|\alpha) \prod_{n=1}^{N} \prod_{m=1}^{M} \left[p(z_n|\theta) \cdot p(w_m|z_n, \gamma_\beta)\right]. \tag{3.3}$$

Equation 3.3 combines the probability of the topic distribution $p(\theta|\alpha)$, the probability of the topic assignment $p(z_n|\theta)$, and the probability of the word given the topic $p(w_m|z_n, \gamma_\beta)$.

Next, to find the marginal distribution of a document, we integrate over the latent variable θ and sum over the latent variable z:

$$p(w|\alpha, \gamma_\beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \prod_{m=1}^{M} \sum_{z_n} p(z_n|\theta) \cdot p(w_m|z_n, \gamma_\beta) \right) d\theta.$$

This step marginalizes out the latent variables to obtain the probability of the observed words w given the hyperparameters α and γ_β.

Finally, we use Bayes' theorem to derive the posterior distribution of the latent variables θ and z given the observed words w and the hyperparameters α and γ_β:

$$p(\theta, z|w, \alpha, \gamma_\beta) = \frac{p(\theta, z, w|\alpha, \gamma_\beta)}{p(w|\alpha, \gamma_\beta)}. \tag{3.4}$$

In Equation 3.4, the numerator is the joint distribution from Equation 3.1, and the denominator is the marginal distribution from Equation 3.2. This posterior distribution is crucial for inferring the hidden structure in the data, but it is generally intractable to compute directly due to the complexity of the integrals involved.

There are different methods that can be applied to estimate the parameters β, θ, and z, such as empirical Bayes, Gibbs Sampling, or Expectation Maximisation (Špeh et al., 2013). These methods help in estimating the topic-word distribution (β), the document-topic distribution (θ), and the topic assignment for each word (z). The application of the empirical Bayes and Expected Maximisation methods can be found in Blei et al. (2003), while the application of the Gibbs Sampling can be found in Porteous et al. (2008).

3.3 BITERM TOPIC MODEL

In contrast to conventional document-centric models, the Biterm Topic Model (BTM) (Yan et al., 2013) is a method to topic modelling that focuses on word co-occurrence patterns over the whole corpus rather than documents. Instead of treating each document as a unit on its own, BTM directly models biterms which are pairs of words that frequently appear together within a short window of text. These biterms capture word relationships outside of specific documents. Word co-occurrences are taken into account by BTM to reveal subtle correlations that document-centric models could overlook. It allows for more precise identification of latent topics. Traditional models often struggle with short texts or sparse data, but BTM's approach mitigates these challenges by leveraging global information from the entire corpus. Unlike other models, BTM embraces the idea that meaningful topic associations can be derived from the relationships between pairs of words. Traditional topic models operate at the document level. They rely on word frequencies within each document to infer topics. However, when dealing with short texts (e.g., tweets, product reviews, or chat messages), individual documents may contain only a few words. As a result, the data can be sparse, lacking sufficient information for accurate topic modelling. The main concept of the BTM is to address the sparsity issue in a single document by learning topics across brief texts based on the aggregated biterms in the entire corpus. In particular, we view the entire corpus as a combination of themes, with each biterm being selected separately from a particular topic. The likelihood that both of the words in a biterm are derived from the topic further captures the probability that the biterm is drawn from a certain topic.

The BTM captures the probability that a biterm is drawn from a specific topic by considering the likelihood that both words in the biterm are generated from the same topic. Suppose α and γ are Dirichlet priors. The BTM models the generation process of biterms within a corpus. This generative process, described in Yan et al. (2013), involves the following steps:

1. For each topic t_l, sample a topic-specific word distribution $\beta_l \sim$ Dirichlet(γ).
2. Sample $\phi \sim$ Dirichlet(α) for the whole collection.
3. For each biterm b in the biterm set B, sample $z \sim$ multinomial(ϕ) and sample $b = (w_i, w_j) \sim$ multinomial(β_l).

Following the process above, the joint probability of a biterm $b = (w_i, w_j)$ can be written as

$$\begin{aligned} p(b) &= \sum_{l}^{T} p(t_l)p(w_i|t_l)p(w_j|t_l) \\ &= \sum_{l}^{T} \phi_l \beta_{i|l} \beta_{j|l} \end{aligned}$$

where ϕ_l represents the probability of topic t_l in the whole collection, $\beta_{i|l}$ is the probability of word w_i occurring given that the topic is t_l and $\beta_{j|l}$ represents the probability of word w_j occurring given that the topic is t_l. The likelihood of all of the biterms in the whole corpus can therefore be written as

$$L(B) = \prod_{i,j} \sum_{l}^{T} \phi_l \beta_{i|l} \beta_{j|l}$$

where B is a set of all of the biterms.

In this model, the focus is on directly modelling word co-occurrence patterns instead of individual words as the units that convey the semantics of topics. It is evident that the co-occurrence of word pairs provides a stronger indication of underlying topics than single words, thereby enhancing topic learning (Yan et al., 2013). Furthermore, by aggregating all biterms from the entire corpus rather than from individual documents, we can fully utilise the extensive global word co-occurrence patterns to more effectively uncover latent topics (Yan et al., 2013).

The topic proportions of documents cannot be derived directly since BTM does not model the document generation process. To infer the topics in a document, there is the assumption that the topic proportions equal the expectation of the topic proportions of biterms generated from the document:

$$p(t_l|d_k) = \sum_b p(t_l|b)p(b|d_k). \tag{3.5}$$

The term $p(t_l|b)$, in Equation 3.5, can be calculated using Bayes' theorem giving the following equation

$$\begin{aligned} p(t_l|b) &= \frac{p(t_l)p(w_i|t_l)p(w_j|t_l)}{\sum_l^T p(t_l)p(w_i|t_l)p(w_j|t_l)} \\ &= \frac{\phi_l \beta_{i|l} \beta_{j|l}}{\sum_t^T \phi_l \beta_{i|l} \beta_{j|l}}. \end{aligned}$$

The term $p(b|d_k)$, in Equation 3.5, is calculated as the empirical distribution of biterms

$$p(b|d_k) = \frac{n_k(b)}{\sum_b n_k(b)}$$

where $n_k(b)$ is the frequency of biterm b occuring in document d_k.

To estimate the parameters ϕ_l and $\beta_{w,l}$ Gibbs Sampling can be used. For more details, see Yan et al. (2013).

3.4 DIRICHLET MULTINOMIAL MIXTURE MODEL

The Dirichlet Multinomial Mixture (DMM) (Yin and Wang, 2014) model stands as a powerful extension of probabilistic topic modelling, designed to uncover latent structures within document corpora (Yin and Wang, 2014). Developed in response to the evolving landscape of text analysis, the DMM model builds upon the foundations laid by models such as LDA (Yin and Wang, 2014).

In both the LDA and BTM models, the topic assignment for each word in a document is sampled from a multinomial distribution. However, in LDA, the word itself is sampled from a multinomial distribution parameterised by the topic-specific word probabilities β_l. In contrast, in the DMM model, the word is assumed to be sampled from a multinomial distribution with parameters corresponding to the assigned topic $z_{d_k,i}$. This allows the model to create a framework that makes it easy to find hidden patterns in textual data.

There are two key assumptions in this generative process (Yin and Wang, 2014). The first is that documents are generated by a mixture model, where each document is drawn from a specific mixture component or topic. The second assumption is that there is a one-to-one correspondence between mixture components and topics. When generating a document d_k, the DMM first selects a topic t according to the mixture weights $p(z = l)$. The document d_k is then generated by the selected topic from the distribution $p(d_k|z = l)$.

The likelihood of document d_k can be characterised by the sum of the total probability over all topics T:

$$p(d_k) = \sum_{l=1}^{T} p(d_k|z = l)p(z = l)$$

where T is the number of topics. To define $p(z = l)$, DMM makes the Naïve Bayes assumption: the words in a document are generated independently when the document's cluster label t_l is known, and the probability of a word is independent of its position within the document. Consequently, the probability of document d_k generated by topic t_l can be derived as:

$$p(d_k|z = l) = \prod_{w \in d_k} p(w|z = l)$$

Each topic is assumed to be a multinomial distribution over words, such that $p(w|z = l) = p(w|z = l, \beta) = \beta_{l,w}$, where $w = 1, 2, \ldots, V$ and $\sum_w \beta_{l,w} = 1$.

The DMM assumes a Dirichlet distribution as the prior for each topic, such that $p(\beta|\gamma_\beta) = \text{Dirichlet}(\beta_l|\gamma_\beta)$. Additionally, it assumes that the weight of each topic is sampled from a multinomial distribution such that $p(z = l) = p(z = l|\theta) = \theta_l$, where $l = 1, 2, \ldots, T$ and $\sum_l \theta_l = 1$. A Dirichlet prior is also assumed for this multinomial distribution, such that $p(\theta|\alpha) = \text{Dirichlet}(\theta|\alpha)$.

In summary, the generative process of the DMM model involves the following steps:

1. For each topic t_l, $l = 1, 2, \ldots, T$, sample $\beta_l \sim$ Dirichlet(γ_β), where β_l represents the topic-word distribution for topic t_l. The Dirichlet distribution serves as a prior over the multinomial distributions governing the topic-word distributions.

2. For each document d_k, $k = 1, 2, \ldots, D$, sample $\theta_{dk} \sim$ Dirichlet(α). The parameter α controls the sparsity of the document-topic distributions.

3. For each document d_k, sample a topic assignment $z_k \sim$ multinomial(θ_k). z_k represents the dominant topic for the entire document.

4. For each word $w_{i,k}$ in document d_k, given the topic assignment z_k, sample $w_{i,k}$ from the multinomial distribution parameterised by $\beta_{z_k, w_{i,k}}$.

The Gibbs Sampling method is used to estimate z_k. Further details can be found in Yin and Wang (2014).

3.5 CONCLUSION

This chapter discussed different topic models, specifically the Latent Dirichlet Allocation (LDA), Biterm Topic Model and the Dirichlet Multinomial Mixture (DMM) topic models. As mentioned before, the LDA topic model is a popular choice for aspect based sentiment analysis. However, it was found that it is better suited for longer pieces of texts, compared to shorter texts (Mazarura et al., 2020; Egger and Yu, 2022). This is because LDA identifies topics by analysing the distribution of words across documents. Longer texts provide more context, allowing LDA to discover coherent topics based on word co-occurrence patterns. In shorter texts, the limited context may lead to less reliable topic assignments due to sparse word occurrences. In comparison to the LDA, the BTM model focuses on modelling the co-occurrence patterns of word pairs within documents. This means that the model captures pairwise relationships directly because the final probability distribution incorporates biterm distributions. On the other hand, the DMM model modifies the LDA by incorporating a Dirichlet-multinomial distribution for topic-word probabilities. A key characteristic of the DMM model is that it assumes each document is associated with a single topic, which is in contrast to LDA, where a document can be a mixture of multiple topics. This assumption makes DMM particularly well-suited for short texts, where it is reasonable to assume that a single topic dominates the document.

With the focus on the aspect-based sentiment analysis, the next chapter will discuss sentiment analysis methods.

CHAPTER FOUR

SENTIMENT ANALYSIS

4.1 INTRODUCTION

Sentiment analysis, also known as opinion mining (Saad and Saberi, 2017), is a Natural Language Processing (NLP) technique that focuses on identifying and extracting sentiments or emotions expressed in text data. It helps determine whether the sentiment behind a piece of text is positive, negative, or neutral (Wankhade et al., 2022). It applies text analysis and computational techniques to automate the extraction or classification of sentiments from texts (Hussein, 2018). Analysis of these sentiments has spread across many fields such as consumer information, marketing, books, websites and social media (Wankhade et al., 2022).

There are two main approaches used in sentiment analysis, namely, machine learning and lexicon-based approaches (Medhat et al., 2014; Birjali et al., 2021; Wankhade et al., 2022). Machine learning algorithms can be grouped into two main categories: supervised learning and unsupervised learning. Supervised learning uses a labelled dataset, where each document is annotated with the corresponding sentiment, to train the algorithm. In contrast, unsupervised learning methods do not rely on labelled data; instead, they identify patterns and structures within the data to infer sentiments without explicit labels. Some of the supervised learning methods that have been successfully applied in sentiment analysis are Multinomial Naïve Bayes, Support Vector Machines and logistic regression (Ahmad et al., 2017). Out of these methods, Multinomial Naïve Bayes is one of the simplest and most common methods to understand and apply. This approach will therefore be discussed in more detail in this chapter.

On the other hand, lexicon-based approaches make use of sentiment lexicons, which are collections of known and pre-compiled sentiment words. Lexicon-based approaches are divided into dictionary-based approaches and corpus-based approaches (Medhat et al., 2014), where dictionary-based approaches are more commonly used. Dictionary-based lexicon approaches rely on predefined sentiment lexicons or dictionaries (Sadia et al., 2018). These lexicons contain words and phrases labelled with their associated sentiment polarity (positive, negative, or neutral). Opinion words are gathered from online dictionaries and other resources to form these lexicons, which include hierarchies, antonyms, and synonyms to determine word sentiments. Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto and Gilbert, 2014) is a popular dictionary-based lexicon used for sentiment analysis and will also be discussed in more detail in the sections that follow.

4.2 MULTINOMIAL NAÏVE BAYES

Multinomial Naïve Bayes (Kibriya et al., 2005) is widely popular and efficient in machine learning. It relies on Bayes' theorem (Rish et al., 2001) and is commonly used in text classification tasks involving discrete data, such as word counts in documents. To grasp Multinomial Naïve Bayes, it is helpful to first understand how the basic Naïve Bayes algorithm works. In this classification model, the posterior probability of a class is computed based on the word distribution within the document (Medhat et al., 2014). The technique uses a bag-of-words feature extraction, disregarding word positions, and leverages Bayes' Theorem to predict the probability that a given feature set corresponds to a specific label.

Bayes theorem, developed by Thomas Bayes (Bayes, 1763), determines the likelihood that an event will occur based on knowledge of the event's prior circumstances, namely:

$$P(s|w_i) = \frac{P(s)P(w_i|s)}{P(w_i)}$$

where $P(s|w_i)$ is the probability of a class s (positive, neutral or negative) given a word w_i, $P(s)$ is the probability of a class occuring, $P(w_i|s)$ is the likelihood of the i^{th} word w_i given a certain class s occurs and $P(w_i)$ is the probability of generating the i^{th} word and it is independent of s. Using this information, Multinomial Naïve Bayes takes the following approach (Abbas et al., 2019):

1. Calculate the prior probabilities of the sentiment classes.

$$p(s) = \frac{N_s}{D}$$

where N_s is the number of documents in class s and D is the total number of documents in the corpus.

2. For each word w_i given a sentiment class s, estimate the conditional probability:

$$p(w_i \mid s) = \frac{N_{w_i,s} + \alpha}{N_s + |V|}$$

where $N_{w_i,s}$ is the number of times word w_i appears in sentiment class s, α is the smoothing parameter (usually set to 1 for Laplace smoothing (Field, 1988)), and $|V|$ is the number of unique words in the corpus, which is the number of words in the vocabulary.

3. Use the Naïve Bayes assumption that states that words are conditionally independent given the sentiment class. This simplifies the joint probability of the words given the sentiment class.

$$p(w_1, w_2, \ldots, w_n \mid s) = \prod_{i=1}^{n} p(w_i \mid s).$$

4. Compute the log probability of each word in document d_k, $w_{i,k}$, given class s,

$$\log p(w_{i,k} \mid s) = \log \left(\prod_{i=1}^{n} p(w_i \mid s) \right) = \sum_{i=1}^{n} \log p(w_i \mid s).$$

Also, calculate the log posterior probability.

$$\log p(s \mid w_{i,k}) = \log p(w_{i,k} \mid s) + \log p(s).$$

5. For the given word $w_{i,k}$, choose the class $\hat{s}$ that maximises the log posterior probability.

$$\hat{s} = \arg\max_{s} \Big(\log p(w_{i,k} \mid s) + \log p(s) \Big)$$

The application of the Multinomial Naïve Bayes will be demonstrated with an example. Consider the educational reviews in Table 4.1:

Table 4.1: An example of student feedback with the sentiments included where 1 indicates a positive sentiment and 0 a negative sentiment.

Review	Sentiment
Great course, I learned a lot!	1
Very boring, I didn't like it.	0
The teacher was excellent.	1
I found the material very useful.	1
The lectures were too long and dull.	0
The class was very engaging.	1
Not helpful at all.	0
The assignments were fun and informative.	1
Too difficult to follow.	0
I enjoyed the interactive sessions.	1

Multinomial Naïve Bayes can be used to classify a new, unseen review. Firstly, the training set consists of the labelled reviews seen in Table 4.1, with each review categorised as either positive or negative. Applying the steps given earlier to Table 4.1 we get the following:

1. The sentiment classes can be divided into 2 groups, namely, positive and negative sentiments. The probability of a positive sentiment occurring is calculated as:

$$p(\text{positive}) = \frac{N_{\text{positive}}}{D} = \frac{6}{10} = 0.6$$

The same can be done for the probability of a negative sentiment occuring:

$$p(\text{negative}) = \frac{N_{\text{negative}}}{D} = \frac{4}{10} = 0.4$$

2. For the sake of demonstration, only the sentence "The course was interesting and informative" will be considered for the remaining steps.

   This step calculates the probability of a word occurring given a sentiment class. Table 4.2 displays the word counts for each of the words that are classified as either a positive or a negative sentence. These word counts are used in the probability calculations. The probability of the word *the* occurring given a positive sentiment can be calculated as follows:

$$p(\text{the} \mid \text{positive}) = \frac{N_{w_i,s} + \alpha}{N_s + |V|} = \frac{5+1}{28+37} = 0.0923$$

   where $|V| = 37$ means there are 37 unique words in Table 4.1.

   The rest of the words in the sentence follow the same logic. Table 4.3 shows the probabilities for all of the words in the selected sentence given the sentiments.

Table 4.2: Words of the example sentence, "The course was interesting and informative", with the word counts from the positive and negative sentences.

Word	Word count in the positive sentences	Word count in the negative sentences
The	5	1
course	1	0
was	2	0
interesting	1	0
and	1	1
informative	1	0

   There are 28 words used in all of the positive reviews, and 20 words used throughout the negative reviews, with the vocabulary being 37 words. Lastly, the conditional probabilities are calculated. Suppose again we use the sentence "The course was interesting and informative". The conditional probabilities are calculated and shown in Table 4.3.

Table 4.3: Probabilities of the words from the example sentence,“The course was interesting and informative”, given it is from a positive or negative class.

Word	Probability (Positive)	Probability (Negative)
the	$\frac{5+1}{28+37} = 0.0923$	$\frac{1+1}{20+37} = 0.0350$
course	$\frac{1+1}{28+37} = 0.0308$	$\frac{0+1}{20+37} = 0.0175$
was	$\frac{2+1}{28+37} = 0.0462$	$\frac{0+1}{20+37} = 0.0175$
interesting	$\frac{1+1}{28+37} = 0.0308$	$\frac{0+1}{20+37} = 0.0175$
and	$\frac{1+1}{28+37} = 0.0308$	$\frac{1+1}{20+37} = 0.0350$
informative	$\frac{1+1}{28+37} = 0.0308$	$\frac{0+1}{20+37} = 0.0175$

3. Using the probabilities from Table 4.3 the joint probabilities can be calculated for each sentiment class. The joint probability of the words in the sentence given the positive sentiment class is calculated as

$$\begin{aligned}
&p(\text{the, course, was, interesting, and, informative} \mid \text{positive}) \\
&\quad = 0.0923 \times 0.0308 \times 0.0462 \times 0.0308 \times 0.0308 \times 0.0308 \\
&\quad = 3.84 \times 10^{-9}.
\end{aligned}$$

The joint probability of the words in the sentence given the negative sentiment class is calculated in the same way having

$$p(\text{the, course, was, interesting, and, informative}|\text{negative}) = 1.15 \times 10^{-10}$$

4. The log probability for the positive sentiment class is calculated as

$$\log p(w_{i,k} \mid s) = \log\left(\prod_{i=1}^{n} p(w_i \mid s)\right) = \log(3.84 \times 10^{-9}) = -19.3833,$$

and the log probability for the negative sentiment class is calculated as

$$\log p(w_{i,k} \mid s) = \log\left(\prod_{i=1}^{n} p(w_i \mid s)\right) = \log(1.15 \times 10^{-10}) = -22.8720.$$

The log posterior probability of the positive class is

$$\log p(s \mid w_{i,k}) = \log p(w_{i,d} \mid s) + \log p(s) = -19.3833 - 0.5108 = -19.8941,$$

The log posterior probability of the negative class is

$$\log p(s \mid w_{i,k}) = \log p(w_{i,d} \mid s) + \log p(s) = -22.8720 - 0.9163 = -23.7883.$$

5. Lastly, the biggest log posterior probability is chosen as the predicted sentiment class. Since $-19.8941 > -23.7883$, the model predicts that the sentence has a positive sentiment.

4.3 VADER LEXICON

Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto and Gilbert, 2014) is a lexicon-based algorithm with a Python package widely used for sentiment analysis. It assigns sentiment scores to individual words or phrases based on their emotional tone. The VADER lexicon contains predefined sentiment terms, and each word is tagged with its corresponding polarity (positive, negative, or neutral) (Hutto and Gilbert, 2014). For example, *happy* has a polarity score of 0.5719 and *angry* has a polarity score of -0.5106. The VADER lexicon was created using a combination of qualitative and quantitative methods. Initially, the creators compiled a list of lexical features from existing sentiment word-banks. They then used crowdsourcing to have human raters evaluate the sentiment intensity of each word or phrase, ensuring accuracy and relevance. This list was empirically validated against social media texts, and rule-based adjustments were made to account for grammatical and syntactical conventions. By summing up token scores within a sentence, VADER determines the overall sentiment. Positive scores contribute to a positive attitude, while negative scores indicate negativity. VADER is particularly effective for short and informal texts, such as social media posts, customer reviews, and educational feedback (Hutto and Gilbert, 2014). VADER analyses a piece of text to see if any of the words from the text are present in the VADER lexicon. It can find the polarity indices using the polarity scores function, which returns the metric values of the negative, neutral, and positive sentiments for a given sentence. The compound score, as seen in Equation 4.1, is a metric that calculates the sum of all the lexicon ratings, which have been normalised between -1 and +1, where -1 indicates the most extreme negative and

+1 indicates the most extreme positive.

To better understand how VADER processes text, we shall examine the steps involved in its sentiment analysis:

1. Tokenisation: Split the input text T into individual words or phrases $W = \{w_1, w_2, \ldots, w_n\}$.

2. For each word w_i in W:

   - Check if w_i exists in the VADER lexicon L.
   - If $w_i \in L$, retrieve its sentiment score o_i, where $o_i \in [-1, 1]$.

3. Calculate the compound score O_{comp}, which is the normalised sum of all sentiment scores o_i:

$$O_{\text{comp}} = \frac{\sum_{i=1}^{n} o_i}{\sqrt{\sum_{i=1}^{n} o_i^2 + \epsilon}} \tag{4.1}$$

   where ϵ is a small positive constant to prevent division by zero and $-1 < O_{\text{comp}} < 1$.

It is useful to set the standardised thresholds for classifying sentences as positive, neutral or negative. The typical threshold values are given below:

- Positive Sentiment: $O_{\text{comp}} \geq 0.05$
- Neutral Sentiment: $-0.05 < O_{\text{comp}} < 0.05$
- Negative Sentiment: $O_{\text{comp}} \leq -0.05$

These specific thresholds were validated to maximise the F_1-score[1] in comparisons with human ratings (Hutto and Gilbert, 2014).

Consider the example from Table 4.1, and specifically, the sentence "The course was interesting and informative". Using VADER to predict the sentiment of the sentence, the first element to consider is the individual scores for the words. From the VADER lexicon, the scores for the individual words are seen in Table 4.4.

[1]The F_1 score is a metric used to evaluate how well a model correctly identifies positive cases while minimising mistakes.

Table 4.4: Individual words from the sentence, “The course was interesting and informative”, with their corresponding sentiment scores from the VADER lexicon.

Word	Score
the	0
course	0
was	0
interesting	0.4019
and	0
informative	0

The words *the*, *course*, *was* and *informative* have a sentiment score of 0 according to VADER, as they do not carry an emotional tone based on its lexicon.

The compound score for this sentence is

$$O_{\text{comp}} = \frac{0+0+0+0.4019+0+0}{\sqrt{(0.4019)^2+1}} = 0.37$$

The compound score for this sentence is 0.37, which is greater than 0.05 and therefore the sentence is classified as positive.

4.4 CONCLUSION

Multinomial Naïve Bayes offers adaptability as it can learn from data, making it suitable for various domains and capable of improving with more training data. It provides probabilistic outputs, giving confidence levels in predictions, and performs effectively with large datasets. However, it requires a labelled dataset for training, which can be time-consuming and resource-intensive, and it is computationally intensive, especially with large datasets. The method also demands careful feature selection and pre-processing to achieve optimal performance. On the other hand, the VADER lexicon is easy to use, requiring no extensive training data, and is computationally efficient, making it ideal for real-time applications. It excels in analysing social media text, effectively handling slang, emojis, and acronyms, and uses a predefined lexicon and rules that can be easily understood and modified (Hutto and Gilbert, 2014). However, its performance is limited by the comprehensiveness of its lexicon, and it struggles to capture new or domain-specific terms accurately. VADER is less adaptable as it does not learn from new data and can struggle with understanding context, sarcasm, or nuanced sentiments (Sharma et al., 2023).

Multinomial Naïve Bayes is suitable when you have access to a substantial labelled dataset, need adaptability and continuous learning, require sentiment scores, or deal with a wide variety of text inputs needing a flexible model (Kamath et al., 2018). The VADER lexicon is preferred when a quick, easy-to-implement solution is needed, especially for real-time applications where speed is crucial, when analysing social media text or informal language, or when a large labelled dataset for training is not available (Hutto and Gilbert, 2014).

Both of the methods will be applied to our educational data to determine which performs better. The methodology for aspect-based sentiment analysis will be discussed in the next chapter.

CHAPTER FIVE

PROPOSED METHODOLOGY FOR ASPECT-BASED SENTIMENT ANALYSIS

5.1 INTRODUCTION

This chapter presents the proposed methodology for extracting and analysing insights from textual data using a combination of topic models and sentiment analysis techniques. Specifically, we detail the use of Latent Dirichlet Allocation (LDA) (Blei et al., 2003), the Biterm Topic Model (BTM) (Yan et al., 2013), and the Dirichlet Multinomial Mixture model (DMM) (Yin and Wang, 2014) for aspect extraction, enhanced with an automatic text summariser, and complemented by the Multinomial Naiïve Bayes (Kibriya et al., 2005) and VADER lexicon sentiment (Hutto and Gilbert, 2014) analysis methods.

To validate the effectiveness of this methodology, we apply it to a labelled dataset of restaurant reviews. Where the presence of labels allows us to assess the performance of the models in capturing both aspects and associated sentiments. This controlled assessment is critical, as it lays the foundation for future applications on real-world datasets that lack labels.

By illustrating the methodology with restaurant reviews, this chapter demonstrates the ability of the proposed methodology to extract and analyse insights from text, providing a deeper understanding of feedback and associated sentiments. This work establishes the basis for applications to unlabelled data, such as our student feedback.

5.2 PROPOSED METHODOLOGY

As discussed in preceding chapters, aspect-based sentiment analysis consists of two main components: aspect extraction and sentiment analysis. Aspect extraction is performed first, followed by sentiment analysis for the extracted aspects. The proposed methodology builds upon the work of Anoop and Asharaf (2018). In their research, they applied the LDA topic model to extract topics, which were then manually labelled to correspond to specific aspects. They used Multinomial Naïve Bayes to determine the sentiments associated with these aspects.

However, there are limitations to their approach. As previously mentioned, the LDA topic model may not be optimal for short pieces of text, such as reviews, and manually labelling the aspects is a time-consuming process. Additionally, their sentiment analysis approach may not yield the best results for all datasets, since not all datasets are necessarily labelled to be trained. To address the issue concerning short pieces of texts, the methodology explained below incorporates models better suited for short texts, namely the BTM and DMM. Additionally, to address the time-intensive process of manual labelling, the proposed methodology incorporates an automatic text summarisation step. Lastly, the methodology also explores an alternative sentiment analysis algorithm to the Multinomial Naïve Bayes.

1. Aspect extraction is performed by applying a topic model — specifically, LDA, BTM, and DMM are investigated — to the clean and processed documents, d_k , $k = 1, 2, \ldots, D$. The BTM and DMM topic models are suggested since they are short text topic models. These models group words into topics, which correspond to different aspects. The output of the topic model is β, the topic-by-word matrix, where β_l, $l = 1, 2, \ldots, T$.

2. Each document is assigned an aspect based on the topic probabilities from θ_k, the document-by-topic matrix for document d_k. Multiple topics can be allocated to a single document, and documents with the same predominant topic are grouped together. Unlike Anoop and Asharaf (2018), who identify topics manually, we propose a method of automatically extracting the topics using text summarisation. This change makes our approach more scalable as it eliminates the manual component. This approach selects the most frequent noun using

part-of-speech tagging in the text as the label. Nouns are chosen because typical educational aspects, such as '*class*', '*tests*', and '*lecturer*', are usually nouns. All the documents in the group are then assumed to be about the extracted aspects, and these sentences are subsequently used to derive the associated sentiment.

3. After the aspects have been extracted, sentiment analysis is applied to the text. Sentiment prediction is performed on each document, d_k, $k = 1, 2, \ldots, D$. Anoop and Asharaf (2018) used Multinomial Naïve Bayes as their sentiment analysis method, however, we also applied the VADER lexicon to the texts to predict the sentiments and compare the results to see which method produces the best results. The reason for adding VADER is that the student data is unlabelled, and VADER, being an unsupervised method, is well-suited for handling such data. We expect VADER to outperform the Multinomial Naïve Bayes because it is specifically designed to work well with social media text and other informal language, which is often the nature of student feedback.

4. Lastly, an aggregation is done to determine which overall sentiments are associated with each aspect. This is achieved by determining the average of the positive, neutral and negative sentiments for a topic. To summarise the sentiments associated with each aspect, we calculate the average scores for positive, neutral, and negative sentiments within each topic.

The proposed methodology will now be demonstrated using a restaurant review dataset. Since the data is labelled, the results will be evaluated using accuracy as the primary metric.

5.3 DATA

The data used for demonstration is a labelled dataset about restaurant reviews from the SemEval-14 Task 4[1] by Pontiki et al. (2014). This dataset consists of 3041 English phrases extracted from restaurant reviews by Ganu et al. (2009). There were 3693 restaurant reviews, each with an assigned sentiment. From these reviews, 633 had neutral sentiments, 805 were negative, and 2164 were positive. Table 5.1 shows a sample of the dataset consisting of reviews with their aspects and polarities. The average sentence length was 19.94 words. The original dataset was used in four different cases, called Subtasks. To enhance its functionality, Pontiki et al. (2014) introduced

[1]The data can be found by following the link: https://alt.qcri.org/semeval2014/task4/

additional annotations for aspect terms in sentences (Subtask 1), aspect term polarities (Subtask 2), coarse aspect categories (Subtask 3), which refer to broader or more general categories into which aspect terms can be grouped, and overall sentence polarities, and aspect category-specific polarities (Subtask 4). Aspect terms refer to specific elements or features within a text that are being evaluated or discussed, while aspect term polarities indicate the sentiment expressed towards these terms. Additionally, aspect category-specific polarities provide a broader understanding of sentiment across categories of aspects. Given our focus on aspect-based sentiment analysis, we specifically use the Subtask 4 data.

Table 5.1: Sample of restaurant reviews with their respective aspects and polarities.

Sentence	**Aspect**	**Polarity**
But the staff was so horrible to us.	staff	negative
To be completely fair, the only redeeming factor was the food, which was above average, but couldn't make up for all the other deficiencies of Teodora.	food	positive
The food is uniformly exceptional, with a very capable kitchen which will proudly whip up whatever you feel like eating, whether it's on the menu or not.	food, kitchen, menu	positive
This is a consistently great place to dine for lunch or dinner.	lunch, dinner	neutral
Not only was the food outstanding, but the little 'perks' were great.	food, perks	positive
Our agreed favorite is the orrechiete with sausage and chicken; usually the waiters are kind enough to split the dish in half so you get to sample both meats.	orrechiete with sausage and chicken, waiters, meats, dish	positive
In fact, while leaving the place we saw two people looking at the menu, and I couldn't help telling them that the food was horrible.	food, menu	negative
Even though the restaurant was packed, we were seated promptly and even asked for a table upstairs with no problems.	table, seated	positive
If you want Americanized Chinese food with your usual watery, generic white sauce, this is your place.	white sauce, Chinese food	negative
Okay service.	service	neutral
Whem asked, we had to ask more detailed questions so that we knew what the specials were.	specials	neutral
The bread is the soft paratha bread (unlike the plain bread they use in Calcutta), and the stuffing is tandoori styled and very flavorful.	bread, paratha bread, stuffing	positive
I love when restaurants think using fancy expensive ingredients makes the food fine cuisine, even with no idea how to use them.	ingredients, cuisine, food	positive
The food is uniformly exceptional, with a very capable kitchen which will proudly whip up whatever you feel like eating, whether it's on the menu or not.	menu	neutral
I recommend this spot to anyone who enjoys fine cuisine at reasonable prices.	cuisine, prices	positive
The staff was very attentive, the ambience lovely, and the food superb.	staff, ambience, food	positive
The Pizza and wine were excellent—the service too—but what really MADE this place was the backyard dining area.	Pizza, wine, service, backyard dining area	positive
The waiters are sweet, the food is tasty, and the bill is never too large.	waiters, food, bill	positive

5.4 EXPERIMENTAL SETUP

All experiments were conducted on a computer running Windows 10, utilising Python 3.6 and Java 20.0.1. The hardware configuration included a 3.50 GHz quad-core processor and 16 GB RAM. The LDA topic model was implemented using the *gensim* package by Řehůřek and Sojka (2010) in Python. On the other hand, the DMM and BTM models were executed in Java[2] according to the specifications provided by Qiang et al. (2019). The VADER lexicon sentiment analysis was implemented using the *vaderSentiment* package by Hutto and Gilbert (2014) in Python.

The LDA model was applied using 50, 100, 150 and 200 topics. Since the dataset was labelled, the accuracy of the results from the LDA model compared to the original data could be calculated. This is useful because it enables the assessment of the model's performance and helps determine the optimal number of topics for precise aspect extraction. From the accuracy scores of the different number of topics it was found that 100 topics had the highest accuracy. The other parameters for the LDA model, namely α, θ, and γ, were set to default values. Specifically, α and θ were both set to 0.01, indicating that they are symmetric. The parameter γ was set to 0.001. For the BTM model different numbers of topics as well as different parameters were tested and it was found that the following parameters produced the best results: 550 topics, $\alpha = 0.2$, $\beta = 0.1$, 1000 iterations, and 20 top words. Lastly, the parameters of the DMM model that yielded the highest accuracy are 700 topics, $\alpha = 0.1$, $\beta = 0.01$, 1000 iterations, and 20 top words.

The number of topics used for each topic model were chosen based on the best accuracy results. For the LDA model, the most relevant topics for each document are identified by selecting the topic with the highest document-topic probability. In contrast, for the BTM and DMM models, the topics that collectively accounted for 80% of the probability distribution of each document were selected.

In contrast, for the BTM and DMM models, the top topics were selected based on their contribution to the distribution of each topic. Specifically, each document is represented as a probability distribution over topics, where the probabilities indicate the proportion of the document associated with each topic. For each document, topics were ranked in descending order of their probabilities, and the smallest set of topics that collectively accounted for at least 80% of the total probability was selected.

[2]https://www.java.com/en/

5.5 DATA PRE-PROCESSING

Before the data could be applied to any models, a thorough pre-processing phase was conducted. Data pre-processing is a crucial step as it ensures that the dataset is cleaned and refined, minimising the inclusion of unnecessary or irrelevant information (Dasu and Johnson, 2003). This process not only enhances the quality of the data but also prepares it for more effective and meaningful analyses. The following steps were followed to prepare the data:

1. Text lowercasing: All text data were converted to lowercase to ensure uniformity in word representation. For example, *The food was delicious* became *the food was delicious*.

2. Expansion of contractions: To facilitate a more comprehensive analysis, contractions were expanded to their full forms. For example, *The meal wasn't that great* was converted to *The meal was not that great*.

3. Removal of numbers and special characters: Numbers and special characters were removed from the texts to focus on the linguistic content. For example, *I gave the service 5 stars* became *I gave the service stars*.

4. Removal of stopwords: Common stopwords, like *'a', 'the', 'is', 'are'*, were eliminated to reduce noise and improve the relevance of the extracted topics.

5. Lemmatization: This is the process of reducing the different forms of a word to one single form, for example, *'dining'* changes to *'dine'*.

5.6 EVALUATION METHODS

Since the restaurant data was labelled, the following performance metric was used (Sokolova and Lapalme, 2009):

- Accuracy: Accuracy $= \frac{\text{True Positives + True Negatives}}{\text{True Positives + True Negatives + False Positives + False Negatives}}$. This gives the overall correctness of the model.

True positive refers to the number of correctly predicted positive sentiments, true negative is the number of correctly predicted negative sentiments. False negative refers to the number of sentiments that were predicted as negative, but are actually positive, and false positive refers to the number of incorrectly predicted positive sentiments, but are actually negative sentiments.

5.7 RESULTS

After cleaning the restaurant reviews, 2019 reviews and 1288 unique aspects were identified. The LDA, DMM and BTM topic models were applied to the clean data and text summarisation, and noun extraction were used for the aspect extraction and labelling. Lastly, the Multinomial Naïve Bayes and VADER sentiment analysis was applied to the data and the results from the different models were compared.

5.7.1 LDA RESULT

Although there were 1288 aspects, the LDA model was trained with only 100 topics, since it had the highest accuracy in predicting the aspects (as seen in Table 5.2). After the topic modelling and text summarisation, some examples of the extracted topics include: *staff, food, menu, meat, taste, cheese, drink, chef, pizza, restaurant, cocktail, selection, try, service*. While most of these aspects are relevant to restaurants, there are aspects, like *try*, that seem out of place. Although *try* is typically used as a verb in most contexts, it can also function as a noun, as in *The player scored a try*. It is likely that the automatic text summariser interpreted *try* as a noun, even though it was used as a verb in the restaurant reviews.

Table 5.2: Accuracy of the aspect extraction using the LDA model with different topic numbers.

Topics	Overall Accuracy	Top 50 Aspects Accuracy
50 topics	0.210	0.489
100 topics	0.230	0.527
150 topics	0.172	0.393
200 topics	0.132	0.311

Table 5.2 shows the accuracy scores for the different numbers of topics used in the topic modelling. There are two accuracy scores presented: the overall accuracy and the accuracy of the top 50 most common aspects in the original data. From the table we can see that the overall accuracy of the model was 23%. This indicates that the model accurately assigned aspects to 23% of the reviews. Upon further investigation, it was noted that the dataset contained 977 aspects that appeared only once, which made it challenging for the model to predict these infrequent aspects. This suggests that the model may struggle with sparse or rarely occurring aspects, which can lead to reduced

performance. To address this issue and focus on more commonly discussed aspects, we decided to focus on the most frequent aspects. This approach allowed us to highlight the aspects about which the majority of customers were writing about. Detecting the most common aspects enables practitioners and users to focus on the issues that concern the majority, potentially making changes that will have a higher impact.

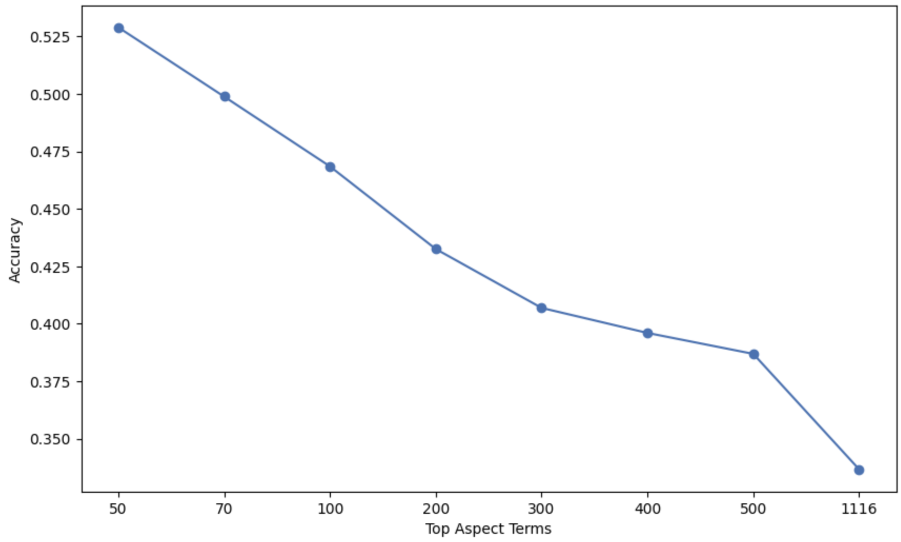

Figure 5.1: Accuracy of the various numbers of most common aspects to occur.

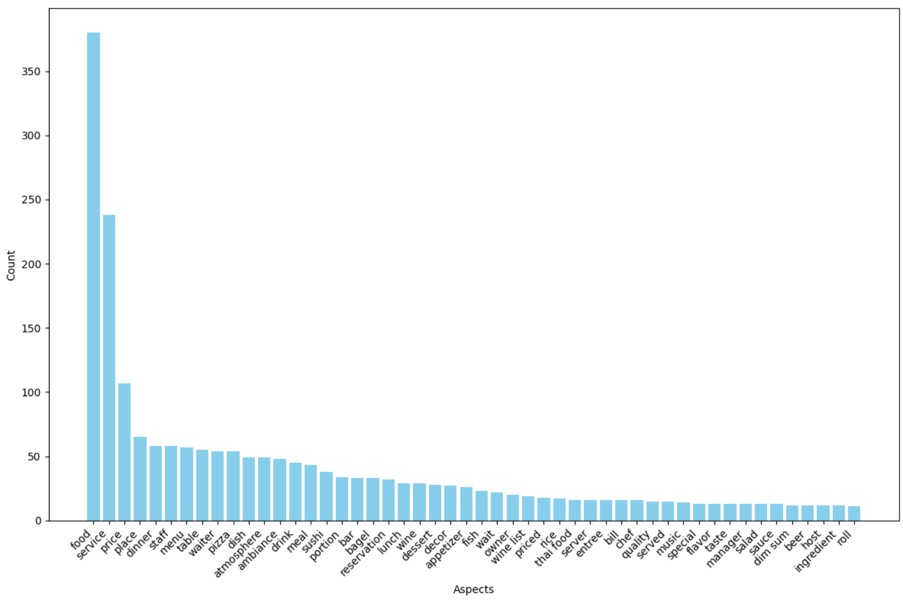

Figure 5.2: Frequency of the 50 most common aspects.

Figure 5.1 visualises the accuracy for different numbers of the most prevalent topics. From Figure 5.1, it is evident that when focusing on the 50 most common aspects, the model presents the highest accuracy. Figure 5.2 visualises the frequency of these 50 most common aspects. The figure shows a steep decline in the frequency of the aspects, which led to the decision to focus on the 50 most common ones. These results indicate that the LDA model is more effective in capturing commonly occurring aspects rather than highly specific ones which are less common.

Table 5.3: A sample of the reviews with the actual and the predicted aspects using the LDA topic model, where matches are in highlighted bold.

Number	Sentence	Actual Aspect	Predicted Aspect
1	But the staff was so horrible to us.	**staff**	**staff**
2	To be completely fair, the only redeeming factor was the food, which was above average, but couldn't make up for all the other deficiencies of Teodora.	**food**	**food**
3	The food is uniformly exceptional, with a very capable kitchen which will proudly whip up whatever you feel like eating, whether it's on the menu or not.	food, kitchen, **menu**	**menu**
4	Not only was the food outstanding, but the little 'perks' were great.	**food**, perks	**food**
5	Our agreed favorite is the orrechiete with sausage and chicken (usually the waiters are kind enough to split the dish in half so you get to sample both meats).	orrechiete with sausage and chicken, waiters, meats, dish	dinner
6	The Bagels have an outstanding taste with a terrific texture, both chewy yet not gummy.	Bagels	taste
7	Nevertheless the food itself is pretty good.	**food**	**food**
8	They did not have mayonnaise, forgot our toast, left out ingredients (ie cheese in an omelet), below hot temperatures and the bacon was so over cooked it crumbled on the plate when you touched it.	toast, mayonnaise, bacon, cheese, ingredients, plate, omelet	wine
9	It took half an hour to get our check, which was perfect since we could sit, have drinks and talk!	**drinks**, check	**drink**
10	Fabulous service, fantastic food, and a chilled out atmosphere and environment.	service, food, atmosphere, environment	food
11	Try the lasagnette appetizer.	lasagnette appetizer	appetizer
12	Though the Spider Roll may look like a challenge to eat, with soft shell crab hanging out of the roll, it is well worth the price you pay for them.	Spider Roll, price, shell crab	roll

Table 5.3 displays a sample of the restaurant reviews with the actual aspects and the predicted aspects from the LDA topic model. The aspects that match are indicated in bold for clarity. The results from the sample show that the LDA model predicts only one aspect per review, as only the top topic was selected for the LDA model. It is also evident that in certain cases it predicts the

correct aspect, as seen in review 1 and 2. However, in other cases, such as in review 8, the model incorrectly predicted the aspect *wine*.

5.7.2 BTM RESULT

The BTM model takes a different approach from the LDA model, particularly in how it adapts the topic count during training. Unlike the LDA model, where the number of topics is fixed from the start, BTM can dynamically adjust the number of topics based on the patterns observed in the data given an initial starting value. This flexibility allows BTM to better capture the variability and nuances in short texts, leading to a more accurate representation of topics as it learns from the data. After experimenting with the number of starting topics, 550 topics were used as an input for the BTM model, since it produced the best accuracy, as seen in Table 5.4. Once the modelling process was completed, 466 topics remained. As the number of distinct topics decreased from 550 to 466 topics, it resulted in broader and more generalised top words and topic labels.

After the text summarisation, the BTM model produced a variety of aspects, from typical restaurant features like *food* and *atmosphere* to surprising aspects like *nothing*, which does not seem to have anything to do with restaurant themes. This could be because *nothing* is not a stopword, so it was not removed during the cleaning process. Additionally, since it is recognised as a noun, it was selected as a label. When the results from the BTM model were compared to the labelled aspects, its overall accuracy was 38%. This measure expresses how well the predicted and true aspects match throughout the dataset. But in order to examine the model's ability to detect prominent topics, we focused our analysis on the 50 most frequently occurring aspects. The model showed a significantly higher accuracy rate of 71%.

Table 5.4: Accuracy of the BTM model with different initial numbers of topics.

Topics	Overall Accuracy	Accuracy Of Top 50 Aspects
100	0.222	0.456
200	0.303	0.582
300	0.346	0.648
500	0.366	0.682
550	0.3845	0.708

Table 5.5: A sample of the reviews with the actual aspects from the dataset versus the aspects predicted from the BTM topic model, where matches are in highlighted bold.

Number	Sentence	Actual Aspect	Predicted Aspect
1	But the staff was so horrible to us.	**staff**	food, **staff**
2	To be completely fair, the only redeeming factor was the food, which was above average, but couldn't make up for all the other deficiencies of Teodora.	**food**	**food**
3	The food is uniformly exceptional, with a very capable kitchen which will proudly whip up whatever you feel like eating, whether it's on the menu or not.	**food**, kitchen, menu	**food**
4	Our agreed favorite is the orrechiete with sausage and chicken (usually the waiters are kind enough to split the dish in half so you get to sample both meats).	orrechiete with sausage and chicken, waiters, meats, **dish**	**dish**
5	The bagels have an outstanding taste with a terrific texture, both chewy yet not gummy.	bagels	sauce, appetizer
6	We ordered the special, grilled branzino, that was so infused with bone, it was difficult to eat.	grilled branzino	portions, experience, chicken, rolls, sushi
7	The wait staff is friendly, and the food has gotten better and better!	wait staff, **food**	**food**, wine,owner, staff, seating, minutes, dim, place
8	It may be a bit packed on weekends, but the vibe is good and it is the best French food you will find in the area.	vibe, french food, packed	food, service
9	Le Pere Pinard has a $15 pre-theater menu that is outstanding.	pre-theater menu	menu
10	What is even better, is that the prices are very affordable as well, and the food is really good.	**prices**, **food**	**food**, **prices**, service, dinner
11	The fish is fresh but the variety of fish is nothing out of ordinary.	**fish**, variety of fish	**fish**, dishes, food, noodles, thai, dumplings, sushi, place
12	Our favorite meal is a pesto pizza, the house salad, and a good bottle of wine.	pesto pizza, house salad, bottle of wine, meal	wine, salad, food, selection, table, pizza, place

Table 5.5 presents a sample of reviews alongside their actual and predicted aspects, with matching aspects highlighted in bold. This comparison allows us to evaluate how well the model captures the essence of the reviews and identifies relevant aspects. It is evident that the model provides concise summaries in some cases, while offering more extensive summaries in others. For instance, review 12 predicted seven aspects, including *selection*, *table*, and *place*, which may not fully align with the review content. Although the model predicts some aspects that do not completely align with the review, it does predict other aspects that are more similar than the model's accuracy might suggest. For example, in review 7, one of the actual aspects is *wait staff*, while the model predicted *staff*. Similarly, in review 9, the actual aspect is *pre-theater menu*, and the model predicted *menu*. Lastly,

in review 12, the model predicted *pizza* instead of *pesto pizza*, *salad* instead of *house salad*, and *wine* instead of *bottle of wine*.

5.7.3 DMM RESULT

Similar to the BTM model, the initial number of topics in the DMM model can differ from the final number of topics, as the DMM model also adapts its topic count based on the data throughout the training process. For example, the DMM model initially began with 700 topics, but after completing the training process, only 644 unique topics remained. Table 5.6 shows the accuracy of the DMM model using different initial topics. From these results it is evident that starting with 700 topics have the best accuracy. After the modelling was done, the groups of texts were summarised. The following aspects are some that were predicted: *saturday, sake, mascarpone, phone, scents, rest, redeeming, chicken, drink, seats*. Although there are aspects that are in line with restaurants, there are aspects seem unrelated to restaurants, like *redeeming*. Following the predictions we calculated the overall accuracy of the model and found that the model is 24% accurate, while the model has a 41% accuracy for the top 50 aspects.

Table 5.6: Accuracy of the DMM model with different initial number of topics.

Topics	Overall Accuracy	Accuracy Of Top 50 Aspects
100	0.159	0.305
300	0.190	0.346
500	0.212	0.366
600	0.234	0.407
700	0.242	0.413

Table 5.7 provides a sample of the results from the DMM modelling method, as well as highlighting the predicted aspects that match the actual aspects. We can see that in most cases only one aspect is predicted for a review. This is very different from the predictions from the BTM model since the BTM model provided multiple aspects.

Table 5.7: A sample of the predictions from the DMM model.

Sentence	Actual Aspect	Predicted Aspect
But the staff was so horrible to us.	**staff**	**staff**
To be completely fair, the only redeeming factor was the food, which was above average, but couldn't make up for all the other deficiencies of Teodora.	food	redeeming
The food is uniformly exceptional, with a very capable kitchen which will proudly whip up whatever you feel like eating, whether it's on the menu or not.	**food**	**food**
Not only was the food outstanding, but the little 'perks' were great.	food	food
Our agreed favorite is the orrechiete with sausage and chicken (usually the waiters are kind enough to split the dish in half so you get to sample both meats).	orrechiete with sausage and chicken	favorite
The bagels have an outstanding taste with a terrific texture, both chewy yet not gummy.	bagels	taste
Nevertheless, the food itself is pretty good.	**food**	**food**
They did not have mayonnaise, forgot our toast, left out ingredients (i.e., cheese in an omelet), served below hot temperatures, and the bacon was so overcooked it crumbled on the plate when you touched it.	toast	family
It took half an hour to get our check, which was perfect since we could sit, have drinks, and talk!	drinks	line
The design and atmosphere are just as good.	design	lunch, service
He has visited Thailand and is quite an expert on the cuisine.	**cuisine**	**cuisine**
The pizza is the best if you like thin-crusted pizza.	pizza	pizza
Wine list selection is good and wine-by-the-glass was generously filled to the top.	wine list selection	wine
With the great variety on the menu, I eat here often and never get bored.	menu	bagels
The menu is very limited - I think we counted 4 or 5 entrees.	**menu**	**menu**

5.7.4 MULTINOMIAL NAÏVE BAYES SENTIMENT ANALYSIS

From the previous results, it is evident that the BTM model has the highest accuracy and therefore it will be the model used for the aspect extraction component. Two subsets were created from the dataset: a training set that contained 75% of the data and a testing set that contained the remaining 25%. The $75\%/25\%$ was used since it is a common training/testing split, providing a good balance between having enough data to train the model effectively and enough data to test its performance accurately. After the model was trained and tested on the original dataset, it correctly predicted 80% of the sentiments. Furthermore, the accuracy of sentiment prediction remained consistent even when the dataset was filtered to include only the top 50 aspects.

Table 5.8 presents the percentage of sentiments for the top ten aspects that were predicted using Multinomial Naïve Bayes.

Table 5.8: Summary of the percentage of each of the sentiment categories for the top ten aspects that were predicted using Multinomial Naïve Bayes.

Aspect	**Positive**	**Neutral**	**Negative**
Dinner	67.37	28.95	3.68
Food	76.75	12.14	11.11
Menu	71.43	17.69	10.88
Pizza	90.53	9.47	0
Place	74	14.8	11.2
Price	74.02	18.63	7.35
Service	74.04	11.92	14.04
Staff	94.44	5.56	0
Table	54.41	27.21	18.38
Waiter	54.84	25.81	19.35

5.7.5 VADER LEXICON SENTIMENT ANALYSIS

The VADER lexicon was also used as an alternative to the supervised Multinomial Naïve Bayes model. Using this approach, it was discovered that 63% of the sentiments in the dataset were correctly predicted by the model. To examine the model's ability to detect prominent topics, we selected the top 50 most frequently occurring aspects. Moreover, the accuracy of sentiment prediction remained consistent even when the dataset was filtered to include only these top 50 aspects.

Table 5.9 presents the percentage of sentiments for the top ten aspects that were predicted, using the VADER lexicon.

Table 5.9: Summary of the percentage of each of the sentiments for the top tenaspects that were predicted using the VADER lexicon.

Aspect	**Positive**	**Neutral**	**Negative**
Dinner	67.79	20.13	12.08
Food	65.99	21.36	12.64
Menu	66.67	19.70	13.64
Pizza	72.60	18.49	8.90
Place	72.42	16.73	10.85
Price	63.86	23.69	12.45
Service	69.26	19.20	11.54
Staff	70.18	20.18	9.65
Table	45.86	36.31	17.83
Waiter	40.23	40.23	19.54

Table 5.10: Summary of the percentage of each of the sentiments for the top tenaspects from the original labelled data.

Aspect	**Positive**	**Neutral**	**Negative**
Dinner	32.73	61.82	5.45
Food	64.99	14.57	20.45
Menu	43.86	28.07	28.07
Pizza	74.42	9.30	16.28
Place	62.5	0	37.5
Price	66.67	7.69	25.64
Service	60.68	7.28	32.04
Staff	78.57	1.79	19.64
Table	29.27	31.71	39.02
Waiter	38.46	11.54	50

Tables 5.10 presents the percentages of the sentiments of the top tenaspects from the original data (Ganu et al., 2009).

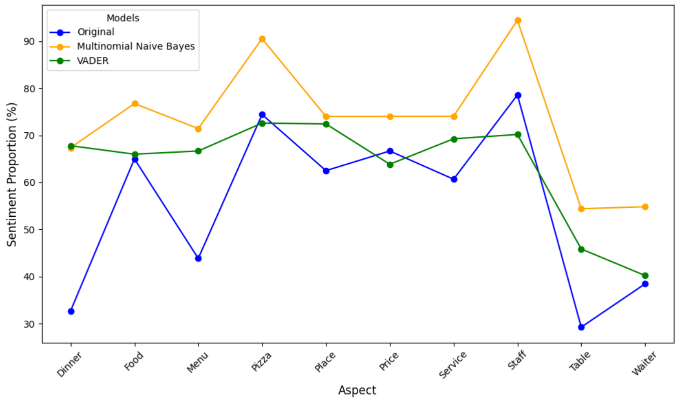

Figure 5.3: A comparison of positive sentiment percentages across the top tenaspects using the original data, the Multinomial Naïve Bayes method and VADER lexicon method.

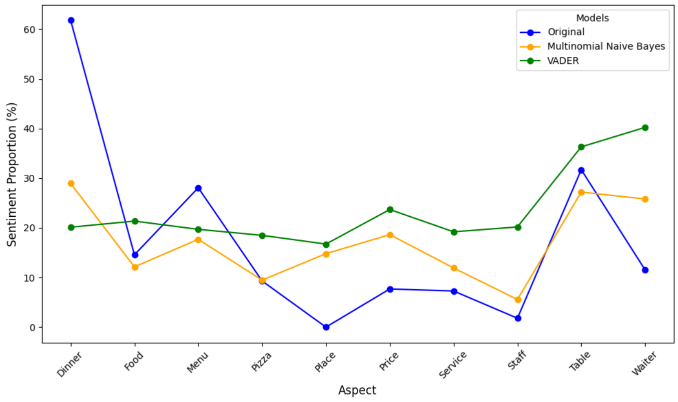

Figure 5.4: A comparison of neutral sentiment percentages across the top tenaspects using the original data, the Multinomial Naïve Bayes method and VADER lexicon method.

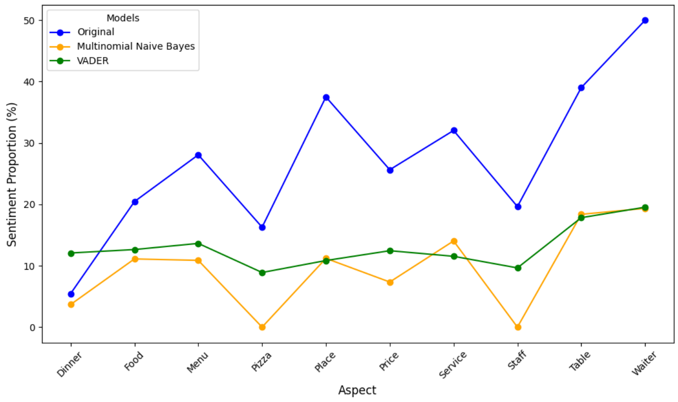

Figure 5.5: A comparison of negative sentiment percentages across the top tenaspects using the original data, the Multinomial Naïve Bayes method and VADER lexicon method.

Figures 5.3, 5.4 and 5.5 compare the percentages of sentiments (positive, neutral, and negative) for the top tenaspects, as derived from the original data, the Multinomial Naïve Bayes method, and the VADER lexicon method, based on Tables 5.8, 5.9 and 5.10. From the figures it is evident that for the first aspect, *dinner*, the original data indicates a predominantly neutral sentiment, while both models predict a more positive sentiment. Although the sentiment percentages for other aspects vary between the original and predicted data, the overall sentiment classifications (positive, neutral, or negative) remain consistent with the original data. Notably, the VADER lexicon predictions for positive sentiments are closer to the original data than those of the Multinomial Naïve Bayes model. However, for neutral and negative classifications, the Multinomial Naïve Bayes model aligns more closely with the original data compared to the VADER lexicon method.

5.8 CONCLUSION

This chapter introduced our proposed new approach for ABSA and was followed by an illustrative example with the use of a labelled restaurant data set. After all of the experiments from the different topic models and the text summarisation, it was evident that the short text topic models (BTM and DMM) produced more accurate predictions in terms of aspect extraction compared to the LDA model, where the BTM model had the highest accuracy. In terms of the sentiment analysis, two methods were applied, the Multinomial Naïve Bayes and the VADER lexicon. From these two methods, the Multinomial Naïve Bayes method showed higher accuracy compared to the VADER lexicon. In the next chapter our methodology will be applied to the student feedback data.

CHAPTER SIX

ASPECT-BASED SENTIMENT ANALYSIS ON STUDENT REVIEW DATA

6.1 INTRODUCTION

This chapter presents a comprehensive analysis of student feedback data, collected from the University of Pretoria. Although this is the first time the data is being analysed, it presents several challenges, particularly due to its unlabelled nature, which prevents direct comparison between predicted and actual aspect or sentiment classifications. Despite these challenges, the analysis offers valuable potential benefits, such as aiding decision-making for management and leadership and identifying areas for improvement in the educational experience.

Building on the results from Chapter 5, which demonstrated that the BTM topic model provided the best aspect extraction performance, this chapter applies the BTM model to the educational dataset. While the previous findings indicate the suitability of the BTM, this chapter includes a detailed explanation to confirm its effectiveness for this specific dataset. In conjunction with the automatic text summarisation step, the BTM model will be used to extract aspects, while two sentiment analysis methods, VADER and Multinomial Naïve Bayes, will be applied to classify sentiment. Together, these methods form the aspect-based sentiment analysis of the student feedback data, with the results discussed in detail in the sections that follow.

This analysis aims to answer key questions, such as:

1. How does sentiment for an aspect vary across different years?
2. How does sentiment for an aspect vary across different academic levels?
3. How does sentiment for an aspect vary across different modules?
4. How does the proposed methodology perform on the education data?

6.2 DATA

The data used in this chapter is an unlabelled student feedback dataset consisting of 10,781 reviews from the Department of Statistics at the University of Pretoria[1], spanning from 2020 to mid-2023. Given that 2020 was significantly impacted by the COVID-19 pandemic, many reviews discuss online learning, technological challenges, and shifts in teaching methods. These factors may have influenced sentiment trends and aspect distributions over time, making it important to consider when comparing pre-pandemic and post-pandemic feedback. The dataset covers four academic levels (first-year to honours) and includes feedback from a total of 35 modules: 16 first-year modules, 7 second-year modules, 7 third-year modules, and 5 honours modules. The average sentence length within the feedback is 12 words. The original data also provides information on the total number of students enrolled in the department, as well as the number of students who gave feedback in each module based on their experiences. Table 6.1 shows a sample of the student feedback data, where the actual module names are replaced with generic module numbers to preserve the confidentiality of the data.

Figure 6.1 shows the boxplots of the response rates from the different modules for the years 2020 up until mid-2023. From the boxplots it is evident that the average response rate over the years is very low. It is also evident that the rates are higher for 2023, where the highest rate was 40%.

[1]The use of this data has been approved under ethics number: NAS124/2023.

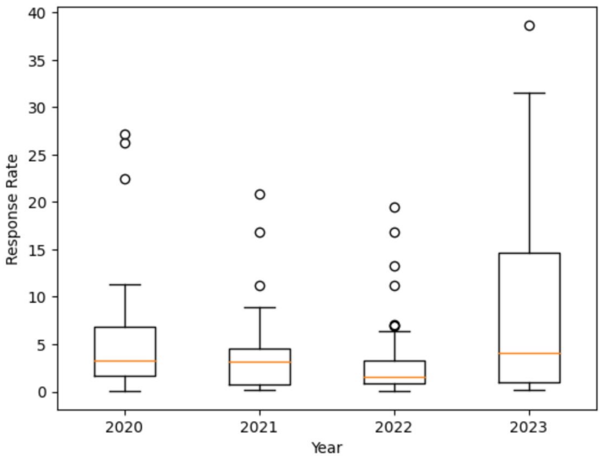

Figure 6.1: Response rate (%) of all of the students for the different Statistics modules from the years 2020 to 2023.

Table 6.1: A sample of the student reviews.

Module	Semester	Year	Review
mod_1	1	2020	A lot of this module was already online before lock down started and so there wasn't too much adjustment that had to take place.
mod_1	1	2020	Able to chop and change between modules to keep brain fresh and prevents boredom
mod_1	1	2020	Able to complete work in my own time and online lectures are very helpful and can be re watched if needed
mod_1	1	2022	The class examples was helpful, but I wish we could get more previous papers that we can practice for preparation for our tests.
mod_1	1	2022	The weekly tutorials helped a lot!
mod_1	1	2022	This module is very tough
mod_13	2	2022	It took a while to adjust but eventually got the hang of it
mod_17	1	2020	I wish the lecturers would maybe explain the work more, rather than just reading of the slides.
mod_28	2	2021	The class tests allow me to ensure I stay up to date with my work.
mod_28	2	2022	The practicals were quite tough.
mod_28	2	2022	Face to face allows me to ask questions if I don't understand something.

Table 6.2: Percentage of reviews from the different academic levels across the different years.

	First year	Second year	Third Year	Honours
2020	73.50%	20.30%	5.75%	0.44%
2021	70.85%	15.87%	12.55%	0.73%
2022	76.68%	8.54%	11.49%	3.29%
2023	82.84%	13.96%	2.64%	0.56%

Table 6.2 shows the percentage of the reviews in the dataset for the different academic levels over the different years. It is evident that the first year students wrote the majority of the reviews each year and that the engagement decreases as the academic levels increase.

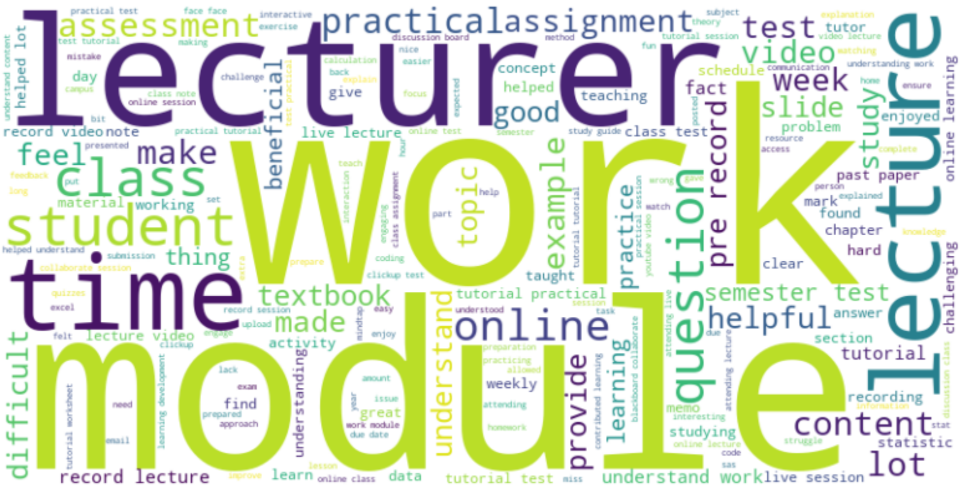

Figure 6.2: Wordcloud of the most prominent words in the comments from the students.

Figure 6.2 visualises a wordcloud from all the comments from the students and from it the words *lecturer, work, module, time* and *lecture* are the most prominent. Table 6.3 presents the response rates for each module from 2020 until mid-2023. While mod_1, a first-year module, had the highest response rates in most years, it was not the highest in 2022. The table also shows that some modules, like mod_7, have a response rate of 0% in certain years. In contrast, mod_35 only has a response rate recorded in 2023. A possible reason for the 0% response rates for certain modules at specific times is that either these modules were not offered during those years, or there is simply no feedback data available for them.

Table 6.3: All of the modules in the Department of Statistics with the percentage of response rates over the years 2020 to mid-2023.

Course	2020	2021	2022	2023	Course	2020	2021	2022	2023
mod_1	19.62%	17.61%	12.73%	29.13%	**mod_19**	2.08%	2.77%	1.03%	0.91%
mod_2	6.00%	6.77%	18.93%	18.45%	**mod_20**	2.44%	2.69%	0.00%	8.74%
mod_3	6.15%	3.78%	1.90%	4.14%	**mod_21**	2.00%	3.02%	1.90%	0.00%
mod_4	3.31%	9.43%	6.45%	14.63%	**mod_22**	2.36%	1.60%	1.49%	0.00%
mod_5	5.67%	5.57%	4.55%	0.00%	**mod_23**	0.44%	0.22%	0.74%	0.00%
mod_6	5.34%	5.24%	5.04%	12.43%	**mod_24**	1.21%	1.89%	0.54%	0.00%
mod_7	1.65%	0.36%	0.00%	0.00%	**mod_25**	2.31%	2.73%	2.19%	0.00%
mod_8	0.04%	0.00%	0.00%	0.00%	**mod_26**	2.23%	4.73%	2.27%	2.59%
mod_9	3.42%	2.66%	10.54%	0.00%	**mod_27**	0.00%	0.58%	1.32%	0.00%
mod_10	2.81%	3.13%	3.18%	0.00%	**mod_28**	0.00%	2.62%	4.71%	0.00%
mod_11	2.92%	3.13%	4.79%	0.00%	**mod_29**	0.00%	0.00%	1.36%	0.00%
mod_12	2.00%	4.51%	3.39%	0.00%	**mod_30**	0.21%	0.00%	0.00%	0.00%
mod_13	0.00%	0.62%	0.58%	0.00%	**mod_31**	0.23%	0.33%	0.12%	0.00%
mod_14	9.90%	3.53%	5.41%	0.00%	**mod_32**	0.00%	0.40%	0.00%	0.00%
mod_15	4.07%	3.78%	0.00%	4.47%	**mod_33**	0.00%	0.00%	1.12%	0.00%
mod_16	0.60%	0.73%	0.37%	0.71%	**mod_34**	0.00%	0.00%	0.04%	0.52%
mod_17	10.05%	5.17%	2.27%	2.33%	**mod_35**	0.00%	0.00%	0.00%	0.13%
mod_18	0.94%	0.40%	1.03%	0.84%					

6.3 EVALUATION

In Section 5.6, the evaluation method that was used for the labelled restaurant data was discussed. However, the student evaluation data used in this chapter is unlabelled and therefore the same evaluation metric will not work. A popular evaluation metric for topic models is the coherence score. The coherence score of a topic is determined by assessing the semantic similarity among words within that topic. The calculation of coherence is expressed as follows (Mimno et al., 2011):

$$\text{coherence}(t_l) = \sum_{(w_i, w_j) \in t_l} \log \frac{D(w_i, w_j) + 1}{D(w_j)}.$$

In this formula, w_i represents the i^{th} word in topic t_l, $D(w_i, w_j) \geq 0$ denotes the number of documents where words w_i and w_j co-occur and $D(w_j) > 0$ is the number of documents where word w_j occurs. A higher average coherence score is preferable, as it indicates more coherent topics, aligning with the inherent goal of achieving meaningful semantic relationships within topics.

6.4 RESULTS

6.4.1 ASPECT EXTRACTION

Figures 6.3, 6.4 and 6.5 present the coherence scores for different numbers of topics. These coherence scores were determined by applying the topic models 10 times for each topic number. From these figures, it is evident that the LDA topic model has the lowest coherence scores, ranging between 0.375 and 0.56, while the DMM model shows coherence scores between 0.53 and 0.59. The BTM model has the highest coherence scores, ranging from 1 to 1.35. Since a higher coherence score is preferable, the BTM model with 250 topics is the best topic model to use for the student data. The LDA model was also trained with 250 topics, while the DMM model was initially set with 50 topics, as these numbers of topics provided the highest coherence scores.

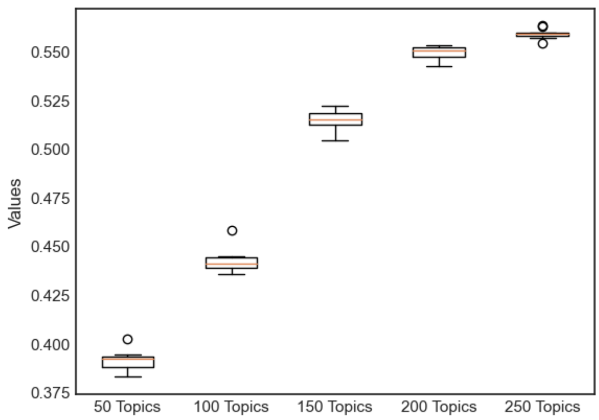

Figure 6.3: Boxplots of the coherence scores for different number of topics for the LDA model

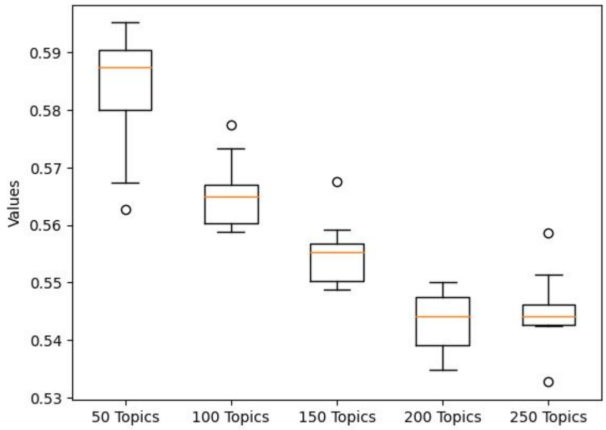

Figure 6.4: Boxplots of the coherence scores for different number of topics for the DMM model

After applying all three topic models, we used the automatic text summarisation to extract and label the aspects. Figures 6.6, 6.7, and 6.8 show the proportion of topics identified by each model, indicating how frequently each topic appears within the dataset. The LDA topic model predicts 91 different topics, however, after the 20^{th} topic the proportion is below 0.5%. Therefore, the proportion of the top 20 topics will be visualised in Figure 6.6. In contrast, the BTM topic model identifies only 17 unique topics, and all are displayed in Figure 6.7. Similarly, the DMM topic model predicts 26 unique topics, which are shown in Figure 6.8. From these graphs, it is clear that the BTM model predicts the fewest topics, suggesting that the BTM identifies more generalised aspects.

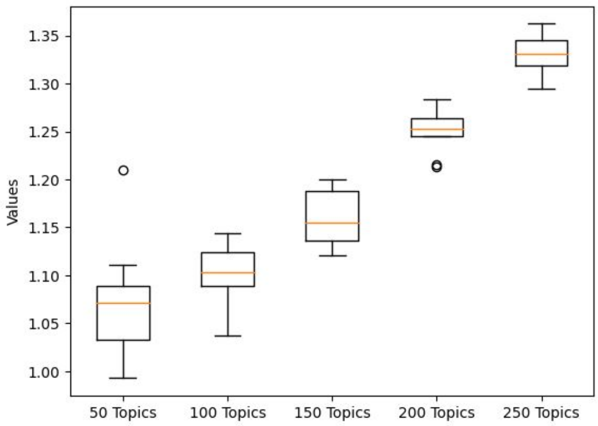

Figure 6.5: Boxplots of the coherence scores for different number of topics for the BTM

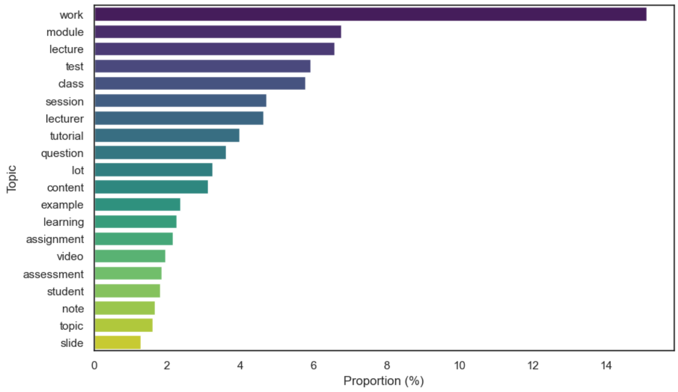

Figure 6.6: Proportion of occurrence for top 20 topics using the LDA model.

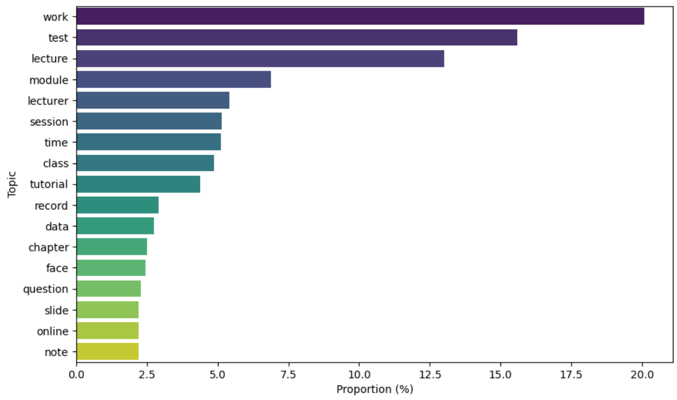

Figure 6.7: Proportion of occurrence for each topic using the BTM model.

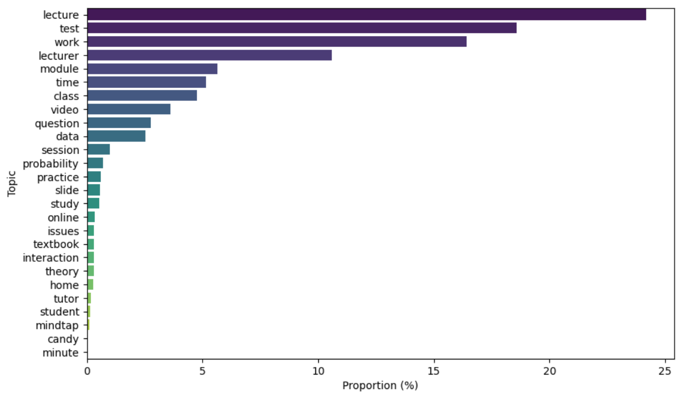

Figure 6.8: Proportion of occurrence for each topic using the DMM model.

Table 6.4: A sample of comments from students and the aspects predicted using the LDA, BTM and DMM topic models, along with the expected aspects based on feedback, where the aspects from the models that match the expected aspects are shown in bold.

No.	Module	Year	Feedback	Expected Aspect	LDA Aspect	BTM Aspect	DMM Aspect
1	mod_5	2020	All past-papers, online assessments and homework are extremely useful and helpful, in understanding content beyond classroom engagement. Very helpful and very grateful for this.	test, assessment, content	paper	test, work, lecture, online	lecture, work, test
2	mod_5	2020	Elaborate explanation with ample opportunity to apply the concepts taught in-class.	class, content	**class**	**class**, work, chapter	time, work, test
3	mod_5	2020	Online tests and homework helped a lot with application. Class helps me with understanding.	test, class	**class**	**test, class**, work, lecture	lecture, work, **test**
4	mod_1	2020	Everything is about your own pace and doing things in your own time, which I prefer, as long as I stick to the weekly work programs. There is a lot of flexibility. I have a lot more time, I'm not wasting 2 hours via the daily transport.	time, class	lot	**time,class** , lecture,work	**time**, lecture
5	mod_1	2020	The tut worksheets help a lot and the excel videos have come in handy.	tutorial, video	lot	test, work	lecture, test
6	mod_2	2020	The lecturers do the best they can to help students.	lecturer	lecture	module, **lecturer**, lecture	**lecturer**, lecture, class
7	mod_7	2021	The practical and tutorials tests really help me understand the content better and helped me prepare for Semester tests. Also attending the Practical and Lecture Blackboard Collaborate sessions really improved my knowledge and understanding of the content.	tutorial, test, practical	**test**	**test**, session, work	lecture, **test**
8	mod_5	2021	The continuous assessment throughout the semester ensured that I stayed up to date and understood the work. The pre-recorded lectures also worked really well.	assessments, lecture	work	test, work, **lecture**	**lecture**, test, work
9	mod_27	2021	The pre-recorded videos were of exceptional quality, as well as the Blackboard Collaborate recordings.	video	**video**	record, work, lecture	lecture, lecturer, module
10	mod_27	2021	Our "lecture videos" were just videos from YouTube that I could've gotten by myself without paying nearly R8000 for this module. The practical videos were the only thing personalized about this module, and even then, the practical lecture was presented on R Studio. We were assessed on SAS. We also didn't get any SAS notes to guide us. That's not even to mention the "YOYO" assignment which is the "you're on your own" assignment that we needed to complete with 0 help from lecturers for 70% of our practical mark. Deeply disappointed in the second quarter of this module.	practical, video, assignment	module	module, test, class, work, lecture	lecture, work
11	mod_4	2021	Watching recorded videos and the tutorials and practicals. The quizzes prepare me for what I don't understand so that I can read more on them to improve my marks.	video, quiz	**video**	test, work, lecture	lecture, test
12	mod_10	2022	Would prefer if this module was online, and they did more work with regards to the coding part of the module. The lectures focus mainly on theory work, which is sometimes grossly unhelpful during the assignments.	class, work	**work**	**class, work**, lecture	lecture, **work**, test
13	mod_12	2022	I preferred the online format as I could focus better compared to face-to-face lectures, where too many students interrupted the lectures.	lecture	**lecture**	module, class, work, **lecture**, face	**lecture**
14	mod_1	2022	Both online and live sessions worked well for me. Recorded lectures also helped improve my understanding.	lecture	record	session, **lecture**	**lecture**, test
15	mod_4	2022	To be honest, the lectures were enough for me to complete the assignments.	lecture	**lecture**	module, work	**lecture**, video, work, test
16	mod_1	2023	He makes the lectures engaging and fun to follow.	class, lecturer	lecture	**lecturer**, work, lecture	**lecturer**, work

Table 6.4 presents a sample of student feedback, including the expected aspects added manually and the aspects predicted by the LDA, BTM and DMM models. The aspects predicted by the models that match the expected aspects are highlighted in bold. Although the LDA model gen-

erates some sensible predictions—for example, entry 9 is correctly associated with *video*—other predictions, such as *lot* in entry 4, are less meaningful. This could be attributed to the noise in the data. The BTM and DMM models tend to make similar predictions, often highlighting words such as *test, module, lecture* and *class*. These words are also emphasised in the word cloud in Figure 6.2, indicating that both models capture aspects that frequently appear in student feedback. As observed in the restaurant data, our aim is to identify the aspects most commonly associated with student feedback. The BTM model stands out for its ability to capture key aspects effectively while achieving higher coherence scores, making it more reliable for further analysis. As a result, the BTM model is selected to present the remainder of the findings.

The following sections will present the results of the aspect-based sentiment analysis using the BTM topic model and the Multinomial Naïve Bayes and VADER lexicon methods.

6.4.2 SENTIMENT ANALYSIS

MULTINOMIAL NAÏVE BAYES

Although Multinomial Naïve Bayes is a supervised machine learning technique, the model was trained using a labelled dataset of over 100,000 course reviews scraped from the Coursera website, available on Kaggle[2]. The reason for this choice is that it is reasonable to assume an educational dataset is expected to contain a consistent vocabulary. The model was applied to the entire dataset, and for example, when applied to first-year students across different years, the aspects of *chapter, class, data, lecture, module, note*, and *online* were analysed. The model predicted all reviews to be negative.

Upon inspection, it was observed that the training dataset was heavily skewed toward negative reviews, which likely influenced the model's predictions. This skewness in the training data created a bias in the learned sentiment proportions, leading the model to predominantly predict negative sentiments. Furthermore, a vocabulary mismatch could have occurred between the Coursera dataset and the student feedback dataset. Words like *module* and *lecture*, which are common in student reviews, might carry different sentiment weights or be underrepresented in the training dataset.

[2] The data can be found using the following link: https://www.kaggle.com/datasets/100k-courseras-course-reviews-dataset

Multinomial Naïve Bayes relies heavily on word frequency, and the overrepresentation of negative sentiment in the training data likely amplified the influence of terms frequently associated with negative sentiment. This could explain why all reviews, including those with varied aspects, were classified as negative. As shown in Table 6.2, first-year students had the highest engagement, suggesting their reviews were more detailed. This richness in vocabulary might have further exposed the model's bias toward negative predictions.

These findings underscore the importance of addressing class imbalance in training data and ensuring alignment with the vocabulary of the target dataset.

VADER

Figures 6.9 - 6.11 show the predicted aspects and their corresponding sentiments over the years 2020 to mid-2023 for first year students. Figures B1 - B3 show the predicted aspects and their corresponding sentiments over the years 2020 to mid-2023 for first year students, which can be found in Appendix I. These academic levels are chosen to visualise since they are the two academic levels with the highest response rates.

From Figure 6.9 the aspect *test* shows a slight increase in positive sentiment from 2020 to 2021. This could be due to the COVID-19 pandemic and the way tests were set up. In 2020, everyone was still getting used to being at home and writing tests online. However, by 2021, lecturers and students had adapted to the online format and were better able to prepare for tests. There is a decrease in sentiment in 2022, which could be attributed to students returning to campus. Most students enrolled in first-year modules in 2022 had not experienced writing tests on campus, which are generally closed-book. Not having the practice or being accustomed to using the internet could negatively affect how students view tests. There is another decrease in positive sentiment in 2023. Since the data from 2023 only covers the first semester, the decline in positive sentiment could be due to students adjusting from high school. Table 6.5 presents a sample of student feedback classified under the *test* aspect. To simplify comparison, the table also includes the expected sentiments for the feedback, which were manually assigned. The results indicate that the model's predictions align with the manually assigned expected sentiments for most of the feedback.

The trends of the aspect *module* in Figure 6.9 are similar to *test*. This similarity may be because students' feelings about modules are influenced by their experiences and outcomes with tests. Similarly, Table A1 presents feedback from first-year students concerning the *module* aspect, which can be found in Appendix I.

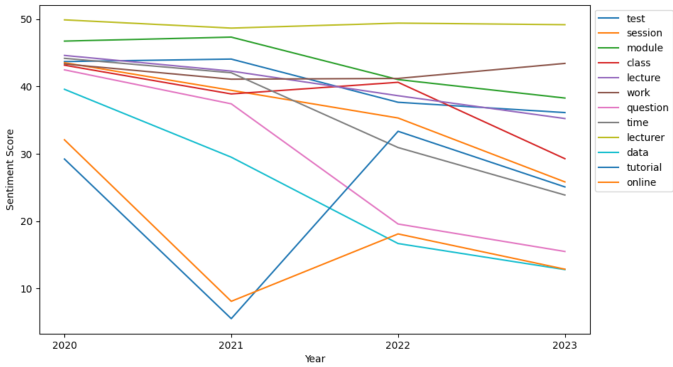

Figure 6.9: Positive sentiments for the aspects for first year students using VADER.

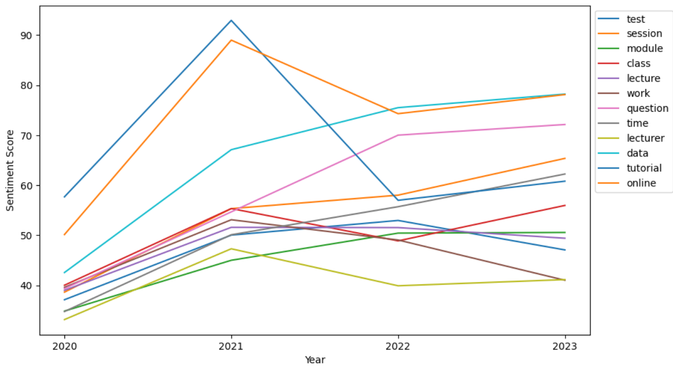

Figure 6.10: Neutral sentiments for the aspects for first year students using VADER.

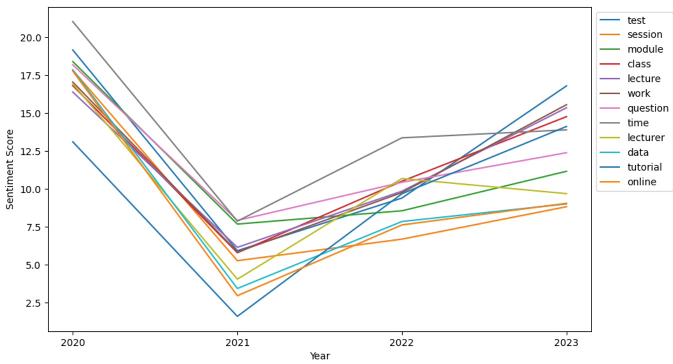

Figure 6.11: Negative sentiments for the aspects for first year students using VADER.

Figure B1, in Appendix I, shows the positive sentiment trends of the aspects associated with second-year students. The aspect *test* starts off with 45% of the feedback being positive in 2020. This positive sentiment in 2020 was influenced by the COVID-19 pandemic. Some students felt they had more resources and tools to help them prepare for tests, while others appreciated the lecturers' efforts to keep everyone up-to-date with the work and test expectations despite the shift to online learning. From 2020 to 2021, there was an increase in positive sentiment. This could be attributed to it being the second year of online learning, with students becoming more familiar with online tests and preparation methods. In 2022, there was a decrease in positive feedback as students returned to campus. The layout of tests likely changed in certain modules, with some shifting from application-based questions online to more theory-based questions in person. The sentiment toward tests further declined in 2023. Some students felt that semester tests were more difficult than class tests, while others faced timing issues, such as having a class immediately before needing to walk to another location for a test — an issue not present with online classes and tests.

A sample of the feedback from the second year students associated with the aspect *test* can be found in Table B1 in Appendix II.

To provide a more detailed analysis, we now shift our focus from the academic year level to the module level. Figures 6.12, 6.13, and 6.14 show the sentiment movement of the aspects associated with module 1. The results are shown for module 1 since it had the highest response rate over most of the years.

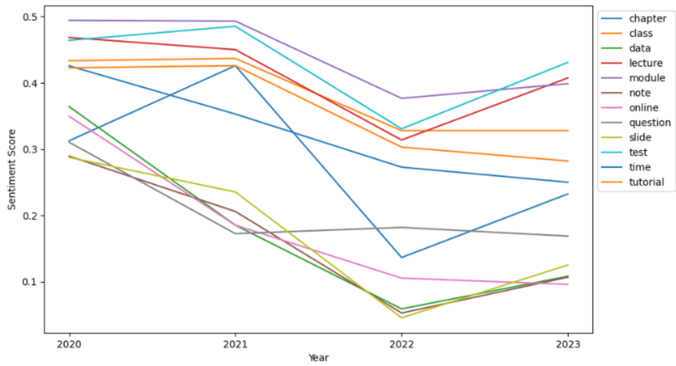

Figure 6.12: Proportion of positive sentiments for the aspects for students taking module 1 using VADER.

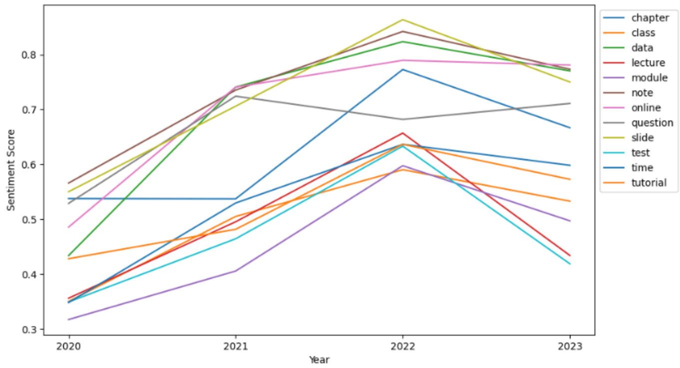

Figure 6.13: Proportion of neutral sentiments for the aspects for students taking module 1 using VADER.

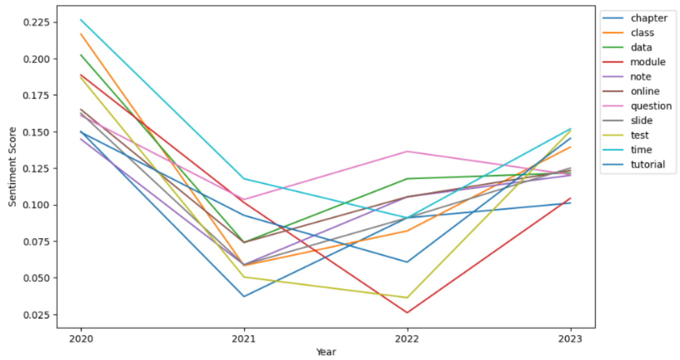

Figure 6.14: Proportion of negative sentiments for the aspects for students taking module 1 using VADER.

At first glance, from Figures 6.12, 6.13, and 6.14, the percentage of negative sentiments for the different aspects is low compared to the positive and neutral sentiments. In particular, the aspect *module* has more positive sentiments than neutral or negative. Table 6.6 presents a sample of student feedback associated with the *module* aspect across different years, along with the manually added expected sentiments. It is evident that most students provided minimal feedback, which could impact the model's performance. However, within the limited data, some students expressed positive opinions, stating that they find the module enjoyable and engaging. The table also demonstrates that, in most cases, the model's predicted sentiments align with the manually added expected sentiments.

Table 6.5: A sample of positive, neutral and negative feedback from first-year students that are associated with the aspect *test*, as well as the expected sentiments for the feedback that have been added manually.

Year	Feedback	Sentiment	Expected Sentiment
2020	The tests we have to write every week to prepare us and tell us where we have to go and look at again and ask questions about.	neutral	neutral
2020	The tests are clear and always on time.	positive	positive
2020	The tests are opened just for one hour and sometimes network disappear and it took me 10 minutes just to get reconnected and everything making it very challenging to complete the test within the hour given. Last week my data got finished and I lost time because I had to buy data.	negative	negative
2021	Tutorial and Practical Tests, Project work assessments and the Class Notes Book as well as the practical guide.	positive	neutral
2021	Having to go to outside sources to get better explanations on topics that were not clearly explained in class. Examples aren't gone through in enough detail in the lecture videos thus leaving you with more questions than answers majority of the time. And having to consult external videos on YouTube to get better understanding, which is quite frustrating. Also not being able to have 1-on-1 consultations is annoying, last year I consulted with a tutor for almost 2 months straight every week and went from 60s to getting 98% in my second semester test. Something that wouldn't have been possible if there weren't any available tutors to consult with.	negative	negative
2022	Classes, tutorials, tutorial tests.	neutral	neutral

Table 6.6: A sample of positive, negative, and neutral feedback from students that are taking module 1 associated with the aspect *module*, as well as the expected sentiments for the feedback that have been added manually.

Year	Feedback	Sentiment	Expected Sentiment
2020	Absolutely everything!	neutral	neutral
2020	I feel like I am taking the extra time to understand the work, and my mental block isn't so big towards the subject anymore. The lecturers and tutors are excellent, and I am really thankful for that!!!!!	positive	positive
2020	The class notes help a lot.	positive	positive
2021	This module was completely useless. I did not learn a single thing from the lectures or this module. I only was able to do well because I did AP Math in high school.	negative	negative
2021	Participation in live sessions and the lecturers doing practical examples with us.	neutral	positive
2022	My lecturer was extremely good!	positive	positive
2022	No comments	negative	neutral
2023	Leaving	neutral	negative
2023	The pre-class assignments. They were really difficult sometimes as we had to teach ourselves the work.	negative	negative
2023	There is nothing not to like. This module is the best, especially because of this lecturer.	positive	positive

6.5 DISCUSSION

From the modelling discussed in the previous sections, several key insights and recommendations can be drawn to help lecturers improve their teaching methods and student support.

Looking at the ABSA results, positive sentiment was found for the aspect of *class*. As a result, the reviews specifically related to the classes were investigated. From these reviews, it was discovered that many students, particularly first, second, and third years, found the pre-recorded videos uploaded for classes very useful, as they could pause, rewind, and review the content at their own pace. Many students also expressed that, after classes resumed on campus, they missed the availability of these pre-recorded videos. Based on these findings, lecturers should consider integrating pre-recorded videos into the curriculum even after returning to in-person classes. This hybrid approach can accommodate different learning styles and support students who may need additional time to grasp complex concepts.

The ABSA results also revealed differences in sentiment regarding the aspect of *test*. While first-year students found pre-tests and post-tests helpful, second and third-year students felt overwhelmed by the additional workload. Students suggested that lecturers consider reducing the number of pre-class and post-class tests for senior students and instead focus on more substantial assessments that integrate these tests' objectives without adding to the workload.

In addition, positive sentiment was found for the aspect of *test* among second-year students. After investigating the reviews, it was noted that these students appreciated continuous assessments instead of tests throughout the semester. These continuous assessments helped them understand the content better and achieve good outcomes. Lecturers should maintain or increase the use of continuous assessments across all years. This approach not only reinforces learning but also provides regular feedback to students, helping them stay on track.

However, there were also negative sentiments toward *test* for second and third-year students. The reviews indicated that students reported errors in assignments and tests, which affected their learning experience. Implementing a peer-review system, where another lecturer reviews assignments and tests before they are distributed to students, can help minimise errors and improve the quality of assessments.

There are clear differences in preferences between junior and senior students regarding teaching methods and assessments. Support and resources should be tailored based on the specific needs of different year groups. For example, providing more structured guidance and frequent assessments for first-year students while offering more flexible and in-depth projects for senior students.

First-year students benefited significantly from assignments and weekly quizzes, which helped them prepare for semester tests. Other modules can adopt similar strategies to enhance student engagement and understanding. Providing examples of questions from previous tests can also help students better prepare for exams.

Some students felt that their questions were not adequately addressed by lecturers or tutors. Improving communication channels between students and faculty can help address this issue. Regular Q&A sessions, office hours, and online forums can provide students with the support they need to clarify confusing concepts.

6.6 CONCLUSION

This chapter introduced the student evaluation data, supplied by the University of Pretoria. The data consists of feedback from students from different academic levels (first-years up to honours students), for years 2020 until mid-2023.

The proposed methodology was applied to the data. The BTM topic model was selected to identify the main aspects in the student feedback. The sentiment analysis using Multinomial Naïve Bayes showed that the assumption that educational content should contain a constant vocabulary is incorrect. The VADER lexicon was found to provide the more insightful results than the Multinomial Naïve Bayes.

Notably, students expressed more positive sentiments towards different aspects in 2020 and 2021, likely due to the shift to online classes during the COVID-19 pandemic, where lecturers ensured sufficient support. However, when students returned to campus, sentiments became more negative. This shift could be attributed to students either becoming accustomed to online classes or facing changes in test formats that included more theoretical questions, which they were not used to.

Given that the data is unlabelled, there is no definitive way to confirm the accuracy of the model predictions. However, after modeling, it is crucial to perform a sense check to ensure that the predictions are reasonable and make sense.

Several key insights and recommendations emerged from the analysis and were discussed in Section 6.5. These findings demonstrate the usefulness of the proposed approach, even in the absence of ground truth labels. The approach highlights key areas for investigation and provides an opportunity to focus on reviews with labels of interest. By generating aspect labels, the model enables lecturers to target specific reviews without the need to read through all of them.

CHAPTER SEVEN

CONCLUSION

This mini-dissertation investigated the application of aspect-based sentiment analysis using topic models on student evaluations in higher education. The inspiration of this mini-dissertation was the paper of Anoop and Asharaf (2018). In their research, they proposed applying the Latent Dirichlet Allocation (LDA) topic model to identify similar words and then used manual labelling to label the aspects. They then applied Multinomial Naïve Bayes as the sentiment analysis.

This mini-dissertation proposed the use of short text topic models, namely the Biterm Topic Model (BTM) and Dirichlet Multinomial Mixture model (DMM), instead of the LDA. The results from the mini-dissertation showed that the short text topic models improved the topic extraction. Specifically, the BTM topic model extracted the most common words in the data. This model was selected because it allows lecturers and the university to identify and address the most frequently mentioned aspects by students.

The next proposal was to apply an automatic text summarisation method that identifies the most frequent noun and uses that as the aspect label, thus alleviating the need for manual labelling. This method helped reduce the modelling time, making the entire process more scalable and less time-consuming.

We also investigated the implementation of two different sentiment analysis methods, the Multinomial Naïve Bayes and VADER lexicon. From these two methods, the results showed that the VADER lexicon method provided more insightful sentiments on the student evaluation data compared to the Multinomial Naïve Bayes method. It was observed that the Multinomial Naïve Bayes model provided sentiments that were heavily skewed toward negative sentiments. This was attributed to the imbalanced dataset used to train the model, which influenced the predictions to overrepresent negative sentiments. Addressing such imbalances in the training data is crucial to obtaining a more accurate representation of sentiments in student evaluations.

Lastly, and most importantly, this mini-dissertation used the student reviews from the Department of Statistics from the University of Pretoria. This is an important contribution since the data has not been analysed in this manner before. Some of the observations from the data showed that most first-year students felt positive about pre-tests and post-tests, while the higher academic levels felt that they already had busy schedules and that these tests take their time away from other assignments. Other insights include integrating pre-recorded videos into the curriculum to accommodate various learning preferences, as students across all years found these videos beneficial for reviewing content at their own pace. Enhancing communication between students and faculty can ensure that student queries are effectively addressed.

The low student response rates impacted the study by limiting the amount of feedback available for analysis, thereby reducing the richness of the dataset for modelling purposes. To address this, lecturers could encourage more active participation in providing feedback, which would help create a more comprehensive dataset. However, it is equally important to account for potential biases in the responses. Students who are facing challenges or struggling may feel a stronger need to voice their concerns, while those who are not struggling might perceive little value in responding, leading to an underrepresentation of their viewpoints. Addressing these imbalances is essential to obtaining a more balanced and representative dataset.

Applying the models on a richer and more representative dataset could improve the results. This could be investigated in future work. Future work also includes addressing class imbalances in the training datasets used for sentiment analysis, particularly for the Multinomial Naïve Bayes method, to mitigate biased predictions and ensure accurate sentiment classifications. Additionally, investigating the use of multi-word labels (Kou et al., 2017) instead of the one-word labels used in this research could enhance the summarisation process and diversify the labels to identify specific aspects. Exploring different automatic text summarisation methods could further improve the analysis, such as Chen and Ren (2021); Sharma and Sharma (2022); Shakil et al. (2024). Another important area for future work is to expand the feedback dataset to enhance the accuracy and robustness of the topic modelling results.

The code for the modelling and the dataset used in this research can be accessed through the following link: https://github.com/JanaduToit/ABSA

REFERENCES

ABBAS, M., MEMON, K. A., JAMALI, A. A., MEMON, S., AND AHMED, A. (2019). Multinomial Naïve Bayes Classification Model for Sentiment Analysis. *International Journal of Computer Science and Network Security*, **19** (3), 62.

AHMAD, M., AFTAB, S., MUHAMMAD, S. S., AND AHMAD, S. (2017). Machine learning techniques for sentiment analysis: A review. *International Journal of Multidisciplinary Sciences and Engineering*, **8** (3), 27.

AL AMRANI, Y., LAZAAR, M., AND EL KADIRI, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, **127**, 511–520.

AL-SMADI, M., TALAFHA, B., AL-AYYOUB, M., AND JARARWEH, Y. (2019). Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. *International Journal of Machine Learning and Cybernetics*, **10**, 2163–2175.

ALBALAWI, R., YEAP, T. H., AND BENYOUCEF, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, **3**, 42.

ALI, F., KWAK, D., KHAN, P., EL-SAPPAGH, S., ALI, A., ULLAH, S., KIM, K. H., AND KWAK, K.-S. (2019). Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems*, **174**, 27–42.

ALI, T., MARC, B., OMAR, B., SOULAIMANE, K., AND LARBI, S. (2021). Exploring destination's negative e-reputation using aspect based sentiment analysis approach: Case of Marrakech destination on TripAdvisor. *Tourism Management Perspectives*, **40**, 100892.

AMPLAYO, R. K., LEE, S., AND SONG, M. (2018). Incorporating product description to sentiment topic models for improved aspect-based sentiment analysis. *Information Sciences*, **454**, 200–215.

ANOOP, V. AND ASHARAF, S. (2018). Aspect-oriented sentiment analysis: A topic modeling-powered approach. *Journal of Intelligent Systems*, **29** (1), 1166–1178.

ATTANASI, E. D. AND COBURN, T. C. (2020). *Random Forest*. Springer International Publishing, Cham, pp. 1–4.

BACCIANELLA, S., ESULI, A., AND SEBASTIANI, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

BAGHERI, A., SARAEE, M., AND DE JONG, F. (2014). ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences. *Journal of Information Science*, **40** (5), 621–636.

BAXTER, R. A. (2010). *Mixture Model*. Springer US, Boston, MA, pp. 680–682.

BAYES, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical Transactions of the Royal Society of London*, (53), 370–418.

BEHDENNA, S., BARIGOU, F., AND BELALEM, G. (2016). Sentiment analysis at document level. *In Smart Trends in Information Technology and Computer Communications: First International Conference, SmartCom 2016, Jaipur, India, August 6–7, 2016, Revised Selected Papers 1*. Springer, pp. 159–168.

BERGER, A., DELLA PIETRA, S. A., AND DELLA PIETRA, V. J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, **22** (1), 39–71.

BHAMARE, B. R., JEYANTHI, P., AND SUBHASHINI, R. (2019). Aspect Level Sentiment Analysis Approaches. *In 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*. IEEE, pp. 1–5.

BHANDARI, S. AND GHOSH, S. (2016). An overview of sentiment analysis: Approaches and applications. *International Journal of Research in Computer Science*, **3**, 4–10.

BHOWMIK, A., NOOR, N. M., MIAH, M. S. U., MAZID-UL-HAQUE, M., AND KARMAKER, D. (2023). A comprehensive dataset for aspect-based sentiment analysis in evaluating teacher performance. *AIUB Journal of Science and Engineering (AJSE)*, **22** (2), 200–213.

BIRJALI, M., KASRI, M., AND BENI-HSSANE, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, **226**, 107134.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3** (Jan), 993–1022.

BONTA, V., KUMARESH, N., AND NAULEGARI, J. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, **8**, 1–6. doi:10.51983/ajcst-2019.8.S2.2037.

BREIMAN, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

BRERETON, R. G. AND LLOYD, G. R. (2010). Support vector machines for classification and regression. *Analyst*, **135** (2), 230–267.

BRODY, S. AND ELHADAD, N. (2010). An unsupervised aspect-sentiment model for online reviews. *In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pp. 804–812.

CASELLA, G. AND GEORGE, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46** (3), 167–174.

CHANG, J., GERRISH, S., WANG, C., BOYD-GRABER, J., AND BLEI, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, **22**.

CHEN, M., TWOREK, J., JUN, H., YUAN, Q., PINTO, H. P. D. O., KAPLAN, J., EDWARDS, H., BURDA, Y., JOSEPH, N., BROCKMAN, G., ET AL. (2021). Evaluating large language models trained on code. *arXiv*.

CHEN, Z., MUKHERJEE, A., AND LIU, B. (2014). Aspect extraction with automated prior knowledge learning. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 347–358.

CHEN, Z. AND REN, J. (2021). Multi-label text classification with latent word-wise label information. *Applied Intelligence*, **51** (2), 966–979.

CHUMAKOV, S., KOVANTSEV, A., AND SURIKOV, A. (2023). Generative approach to Aspect Based Sentiment Analysis with GPT Language Models. *Procedia Computer Science*, **229**, 284–293.

CORTES, C. AND VAPNIK, V. (1995). Support-vector networks. *Machine Learning*, **20**, 273–297.

DASU, T. AND JOHNSON, T. (2003). *Exploratory data mining and data cleaning*. John Wiley & Sons.

DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

EGGER, R. AND YU, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, **7**, 886498.

FAUZI, M. A. (2018). Random Forest Approach for Sentiment Analysis in Indonesian. *Indonesian Journal of Electrical Engineering and Computer Science*, **12**, 46–50.

FIELD, D. A. (1988). Laplacian smoothing and Delaunay Triangulations. *Communications in Applied Numerical Methods*, **4** (6), 709–712.

GANU, G., ELHADAD, N., AND MARIAN, A. (2009). Beyond the stars: Improving rating predictions using review text content. *In WebDB*, volume 9. pp. 1–6.

GROOTENDORST, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*.

GUO, C., DU, Z., AND KOU, X. (2018). Products ranking through aspect-based sentiment analysis of online heterogeneous reviews. *Journal of Systems Science and Systems Engineering*, **27**, 542–558.

Gupta, N. and Agrawal, R. (2020). Chapter 1 - application and techniques of opinion mining. *In Hybrid Computational Intelligence*, Hybrid Computational Intelligence for Pattern Analysis and Understanding. Academic Press, pp. 1–23.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2004). Support Vector Machines for High-Dimensional Data Classification: A Review. *Bioinformatics*, **20** (3), 273–285.

Hajrizi, R. and Nuçi, K. P. (2020). Aspect-based sentiment analysis in education domain. *arXiv*.

Han, H., Zhang, Y., Zhang, J., Yang, J., and Zou, X. (2018). Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias. *PloS One*, **13** (8), e0202523.

Hidayati, N. N. (2023). Improving aspect-based sentiment analysis for hotel reviews with Latent Dirichlet Allocation and Machine Learning Algorithms. *Register*, **9** (2), 144–159.

Hoang, M., Bihorac, O. A., and Rouces, J. (2019). Aspect-based sentiment analysis using bert. *In Proceedings of the 22nd Nordic Conference on Computational Linguistics*. pp. 187–196.

Hoffman, M., Bach, F., and Blei, D. (2010). Online learning for Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems*, **23**.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 50–57.

Hollander, M. and Sethuraman, J. (2001). Nonparametric statistics: Advanced computational methods. *Journal of Statistical Planning and Inference*, **99** (2), 225–238.

Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, **30** (4), 330–338.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 8. pp. 216–225.

JAMADI KHIABANI, P., BASIRI, M. E., AND RASTEGARI, H. (2020). An improved evidence-based aggregation method for sentiment analysis. *Journal of Information Science*, **46** (3), 340–360.

JAYNES, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, **106** (4), 620.

JO, Y. AND OH, A. H. (2011). Aspect and sentiment unification model for online review analysis. *In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. pp. 815–824.

JOACHIMS, T. (2005). Text categorization with support vector machines: Learning with many relevant features. *In Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings*. Springer, pp. 137–142.

JURAFSKY, D. AND MARTIN, J. H. (2014). *Naïve Bayes, Text Classification, and Sentiment Analysis*. Pearson.

KAMATH, C. N., BUKHARI, S. S., AND DENGEL, A. (2018). Comparative study between traditional machine learning and deep learning approaches for text classification. *In Proceedings of the ACM Symposium on Document Engineering 2018*. pp. 1–11.

KASTRATI, Z., IMRAN, A. S., AND KURTI, A. (2020). Weakly supervised framework for aspect-based sentiment analysis on students' reviews of MOOCs. *IEEE Access*, **8**, 106799–106810.

KAUSAR, S., HUAHU, X., AHMAD, W., AND SHABIR, M. Y. (2019). A sentiment polarity categorization technique for online product reviews. *IEEE Access*, **8**, 3594–3605.

KIBRIYA, A. M., FRANK, E., PFAHRINGER, B., AND HOLMES, G. (2005). Multinomial Naïve Bayes for text categorization revisited. *In AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*. Springer, pp. 488–499.

KIRITCHENKO, S., ZHU, X., AND MOHAMMAD, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, **50**, 723–762.

KOU, W., LI, F., AND YE, Z. (2017). Using topic labels for text summarization. *In Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part II 30*. Springer, pp. 448–457.

LAFFERTY, J., MCCALLUM, A., AND PEREIRA, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-01)*, 282–289.

LI, C., DUAN, Y., WANG, H., ZHANG, Z., SUN, A., AND MA, Z. (2017). Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems (TOIS)*, **36** (2), 1–30.

LIN, C. AND HE, Y. (2009). Joint sentiment/topic model for sentiment analysis. *In Proceedings of the 18th ACM Conference on Information and Knowledge Management*. pp. 375–384.

LIU, B., HU, M., AND CHENG, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. *In Proceedings of the 14th International Conference on World Wide Web*. pp. 342–351.

LIU, L., TANG, L., DONG, W., YAO, S., AND ZHOU, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, **5**, 1–22.

LORIA, S. ET AL. (2018). textblob documentation. *Release 0.15*, **2** (8), 269.

MAZARURA, J., DE WAAL, A., AND DE VILLIERS, P. (2020). A gamma-poisson mixture topic model for short text. *Mathematical Problems in Engineering*, **2020**, 1–17.

MEDHAT, W., HASSAN, A., AND KORASHY, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, **5** (4), 1093–1113.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., AND DEAN, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, **26**.

MIMNO, D., WALLACH, H., TALLEY, E., LEENDERS, M., AND MCCALLUM, A. (2011). Optimizing semantic coherence in topic models. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. pp. 262–272.

NAKOV, P., RITTER, A., ROSENTHAL, S., SEBASTIANI, F., AND STOYANOV, V. (2016). Semeval-2016 task 4: Sentiment analysis in twitter. *In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pp. 1–18.

NIGAM, K., MCCALLUM, A. K., THRUN, S., AND MITCHELL, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, **39**, 103–134.

OZYURT, B. AND AKCAYOL, M. A. (2021). A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA. *Expert Systems with Applications*, **168**, 114231.

PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

PANG, B., LEE, L., AND VAITHYANATHAN, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

PATHIK, N. AND SHUKLA, P. (2022). Aspect based sentiment analysis of unlabeled reviews using linguistic rule based LDA. *Journal of Cases on Information Technology (JCIT)*, **24** (3), 1–19.

PHAN, X.-H. AND NGUYEN, C.-T. (2007). Gibbslda++: A c/c++ implementation of latent dirichlet allocation (lda). *Journal of Machine Learning Research*, **8**, 1–48.

PONTIKI, M., GALANIS, D., PAPAGEORGIOU, H., MANANDHAR, S., AND ANDROUTSOPOULOS, I. (2015). SemEval-2015 task 12: Aspect based sentiment analysis. *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, pp. 486–495.

PONTIKI, M., GALANIS, D., PAVLOPOULOS, J., PAPAGEORGIOU, H., ANDROUTSOPOULOS, I., AND MANANDHAR, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis.

In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Association for Computational Linguistics, Dublin, Ireland, pp. 27–35.

PORTEOUS, I., NEWMAN, D., IHLER, A., ASUNCION, A., SMYTH, P., AND WELLING, M. (2008). Fast collapsed gibbs sampling for Latent Dirichlet Allocation. *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 569–577.

QIANG, J., CHEN, P., WANG, T., AND WU, X. (2017). Topic modeling over short texts by incorporating word embeddings. *In Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II 21*. Springer, pp. 363–374.

QIANG, J., ZHENYU, Q., LI, Y., YUAN, Y., AND WU, X. (2019). Short text topic modeling techniques, applications, and performance: A survey. *arXiv*.

QUAN, X., KIT, C., GE, Y., AND PAN, S. J. (2015). Short and sparse text topic modeling via self-aggregation. *In 24th International Joint Conference on Artificial Intelligence, IJCAI 2015*. AAAI Press/International Joint Conferences on Artificial Intelligence, pp. 2270–2276.

RANA, S. AND SINGH, A. (2016). Comparative analysis of sentiment orientation using SVM and Naïve Bayes techniques. *In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*. IEEE, pp. 106–111.

RASCHKA, S. (2014). Naïve bayes and text classification i-introduction and theory. *arXiv*.

ŘEHŮŘEK, R. AND SOJKA, P. (2010). Software framework for topic modelling with large corpora. *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pp. 45–50.

RISH, I. ET AL. (2001). An empirical study of the Naïve Bayes classifier. *In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, volume 3. Citeseer, pp. 41–46.

SAAD, S. AND SABERI, B. (2017). Sentiment analysis or opinion mining: A review. *International Journal on Advanced Science, Engineering and Information Technology*, **7**, 1660.

SADIA, A., KHAN, F., AND BASHIR, F. (2018). An overview of lexicon-based approach for sentiment analysis. *In 2018 3rd International Electrical Engineering Conference (IEEC 2018)*. pp. 1–6.

SAIDI, F., TRABELSI, Z., AND THANGARAJ, E. (2022). A novel framework for semantic classification of cyber terrorist communities on twitter. *Engineering Applications of Artificial Intelligence*, **115**, 105271.

SANCHEZ-MEDINA, J. J. (2024). Sentiment analysis and random forest to classify LLM versus human source applied to Scientific Texts. *arXiv preprint arXiv:2404.08673*.

SCHMITT, M., STEINHEBER, S., SCHREIBER, K., AND ROTH, B. (2018). Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. *arXiv*.

SEBASTIANI, F. AND ESULI, A. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *In Proceedings of the 5th International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA) Genoa, Italy, pp. 417–422.

SHAKIL, H., FAROOQ, A., AND KALITA, J. (2024). Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing*, 128255.

SHAMS, M. AND BARAANI-DASTJERDI, A. (2017). Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction. *Expert Systems with Applications*, **80**, 136–146.

SHARMA, D. K., SINGH, B., AGARWAL, S., PACHAURI, N., ALHUSSAN, A. A., AND ABDALLAH, H. A. (2023). Sarcasm detection over social media platforms using hybrid ensemble model with fuzzy logic. *Electronics*, **12** (4), 937.

SHARMA, G. AND SHARMA, D. (2022). Automatic text summarization methods: A comprehensive review. *SN Computer Science*, **4** (1), 33.

SIVAKUMAR, M. AND REDDY, U. S. (2017). Aspect based sentiment analysis of students opinion using machine learning techniques. *In 2017 International Conference on Inventive Computing and Informatics (ICICI)*. IEEE, pp. 726–731.

SOKOLOVA, M. AND LAPALME, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, **45** (4), 427–437.

SPARCK JONES, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, **28** (1), 11–21.

ŠPEH, J., MUHIC, A., AND RUPNIK, J. (2013). Parameter estimation for the Latent Dirichlet Allocation. *In Proceedings of the Conference on Data Mining and Data Warehouses, Ljubljana, Slovenia*. pp. 1–4.

STEINWART, I. AND CHRISTMANN, A. (2008). *Support Vector Machines*. Springer Science & Business Media.

STINE, R. A. (2019). Sentiment analysis. *Annual Review of Statistics and its Application*, **6**, 287–308.

TANG, F., FU, L., YAO, B., AND XU, W. (2019). Aspect based fine-grained sentiment analysis for online reviews. *Information Sciences*, **488**, 190–204.

TROUSSAS, C., VIRVOU, M., ESPINOSA, K. J., LLAGUNO, K., AND CARO, J. (2013). Sentiment analysis of Facebook statuses using Naïve Bayes classifier for language learning. *In IISA 2013*. IEEE, pp. 1–6.

TURNEY, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 417–424.

VOSKERGIAN, D., BAKIR-GUNGOR, B., AND YOUSEF, M. (2023). TextNetTopics Pro, a topic model-based text classification for short text by integration of semantic and document-topic distribution information. *Frontiers in Genetics*, **14**, 1243874.

VU, L. AND LE, T. (2017). A lexicon-based method for sentiment analysis using social network data. *In Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*. pp. 10–16.

WAHID, J. A., HUSSAIN, S., WANG, H., WU, Z., SHI, L., AND GAO, Y. (2021). Aspect oriented Sentiment classification of COVID-19 twitter data; an enhanced LDA based text analytic approach. *In 2021 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*. IEEE, pp. 271–275.

WALLACH, H. M. (2006). Topic modeling: Beyond bag-of-words. *In Proceedings of the 23rd International Conference on Machine Learning*. pp. 977–984.

WANG, S., CHEN, Z., AND LIU, B. (2016). Mining aspect-specific opinion using a holistic lifelong topic model. *In Proceedings of the 25th International Conference on World Wide Web*. pp. 167–176.

WANKHADE, M., RAO, A. C. S., AND KULKARNI, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, **55** (7), 5731–5780.

WENG, J., LIM, E.-P., JIANG, J., AND HE, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. *In Proceedings of the third ACM International Conference on Web Search and Data Mining*. pp. 261–270.

WICKRAMASINGHE, I. AND KALUTARAGE, H. (2021). Naïve Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, **25** (3), 2277–2293.

WU, X., LI, C., ZHU, Y., AND MIAO, Y. (2020). Short text topic modeling with topic distribution quantization and negative sampling decoder. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1772–1782.

YADAV, K., KUMAR, N., MADDIKUNTA, P. K. R., AND GADEKALLU, T. R. (2021). A comprehensive survey on aspect-based sentiment analysis. *International Journal of Engineering Systems Modelling and Simulation*, **12** (4), 279–290.

YAN, X., GUO, J., LAN, Y., AND CHENG, X. (2013). A biterm topic model for short texts. *In Proceedings of the 22nd international conference on World Wide Web*. pp. 1445–1456.

YIN, J. AND WANG, J. (2014). A Dirichlet multinomial mixture model-based approach for short text clustering. *In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 233–242.

ZHANG, L. AND LIU, B. (2017). *Sentiment Analysis and Opinion Mining*. Springer US, Boston, MA, pp. 1152–1161.

ZHAO, H., PHUNG, D., HUYNH, V., JIN, Y., DU, L., AND BUNTINE, W. (2021). Topic modelling meets deep neural networks: A survey. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 4713–4720.

ZHAO, X., JIANG, J., YAN, H., AND LI, X. (2010). Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 56–65.

]

FIRST YEAR

I DATA

Table A1: A sample of positive, negative and neutral feedback from first year students that are associated with the aspect *module*.

Year	Feedback	Sentiment
2020	I feel out of all my modules the mod_1 team were the best prepared for the online platform and i really enjoy the sessions.	positive
2020	Nothing i like about online learning in this module.	negative
2020	Not in this module.	neutral
2020	As much as some of the online work is helpful, I feel like I am constantly drowning in the amount of work I have to do for this module. I often spend an entire day doing the assignments trying to get ahead and then I check Click up two days later and there are more assignments due in weeks which already have so many assignments for this module already.	positive
2021	All the abovementioned things helped me to learn and develop in this module; more especially live sessions.	neutral
2021	All the materials that were made available to me (tutorials, practicals, pre and post class assignments, excel activities, live sessions etc.) contributed to my learning and development in this module.	positive
2022	It is not beneficial, the only way to pass the module is to attend lab sessions and the Q&A sessions. The classes are so confusing, I walk out of the class feeling like I wasted my time. I understand the module is still new but there should be convention in our this module is run, truly what we do ion class is not helpful. I am really struggling in this module and I feel like I am not taught anything.	negative
2022	I have improved my understanding and work ethic in this module compared to the 1st semester module.	positive
2022	The use of recordings and notes helped me in preparation for the module tests.	neutral
2022	Blackboard live lectures helped me understand a lot about this module and a lot of practicing helped out a lot too.	neutral
2023	What i like least about this module is that is one of the most challanging modules and requires lot of time.	positive
2023	The lecturer could not explain anything and she did not understand what she teaching. I did not like that attendance was taken so strictly in this module as the lectures were not beneficial and therefore a waste of my time, where I could be teaching myself the content.	negative

]

SECOND YEAR

I SENTIMENT GRAPHS

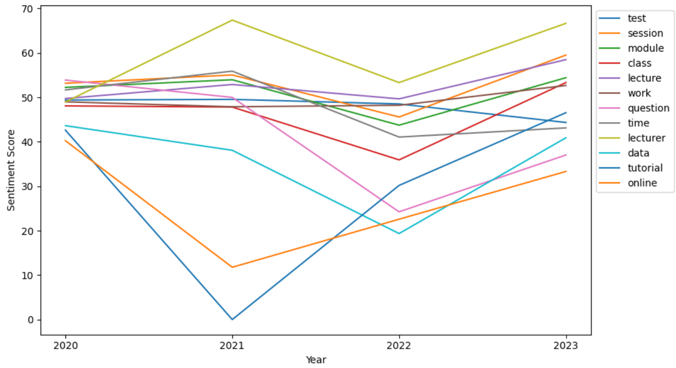

Figure B1: Positive sentiments for the aspects for second year students using VADER.

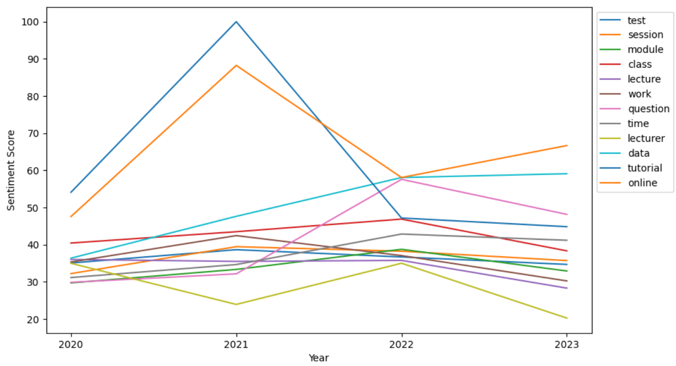

Figure B2: Neutral sentiments for the aspects for second year students using VADER.

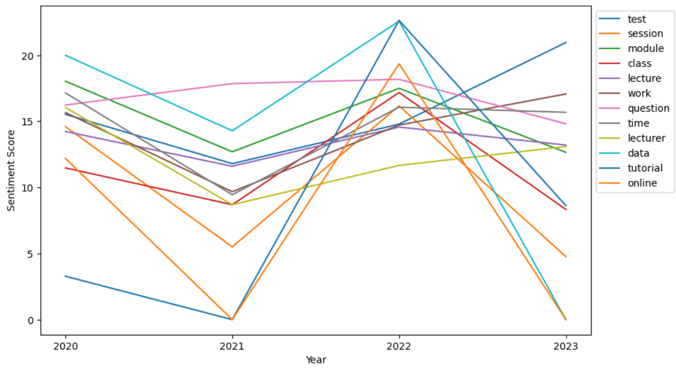

Figure B3: Negative sentiments for the aspects for second year students using VADER.

II SECOND YEAR DATA

Table B1: A sample of positive, neutral and negative feedback from second year students that are associated with the aspect *test*.

Year	Feedback	Sentiment
2020	Homework questions, online tests and practical videos contributed to my learning and development.	neutral
2020	I felt as though both the online theory and practical tests as well as the online practical lectures were especially beneficial when trying to acquire further knowledge. I do however find that the notes are not as comprehensive as i had hoped from last year.	positive
2020	I enjoy that I am forced to look through the work. Since, previously we go to lectures trying hard to pay attention, then go home do assignment, study for an upcoming test and then go to bed without having revised the work you heard in class, so you really don't fully understand the work.	negative
2020	I finally understand the SAS programming expected of us and both lecturers show a great understanding of the work and know how to pass the concepts to us. The examples done in the video lectures and in class earlier this year really helped prepare us for more difficult problems. When I encountered problems with online tests, the lecturers did not hesitate to help me resolve the problem with my test. The lecturers are very approachable.	negative
2021	The practical live lectures are very helpful, l honestly feel as if our lecturers' test what they teach and they give us plenty of practice materials which is very beneficial as it helps us hone in on our skills.	positive
2021	The pre recorded lecture videos and feedback on test.	neutral
2021	Using the homework and memos as well as the test with memos to correct my faults and learn how to do equations I got wrong.	negative
2021	The homework assignments really helped a lot in understanding the theory tests.	neutral
2022	The homework help me test everything I knew the weekly practicals served as small test for the casual practical an the prep videos really help get you ready for the week on what questions you would ask and what you would struggle with.	positive
2022	The homework sheets helped me prepare for semester tests a lot.	neutral
2022	The homework, past papers as well as the online theory and practical tests contributed to my learning and development and helped me prepare for Semester tests.	positive
2023	We did not have a lot of practice material to use for tests.	neutral
2023	Wednesday test and practical session is really bad timing for me and most students' schedule. A lot of us have to make special arrangements to get to campus to attend these session and tests, and it is way more inconvenient than I initially thought. How many times more will we have to be excused early because of semester tests that happen directly after the session? It's just bad timing. I admit writing the test and doing the lecture online is way more preferable than on campus, and I'm overjoyed when Wednesday sessions are online.	negative