UNIVERSIDADE CATÓLICA PORTUGUESA

Using Sentiment Analysis to predict Amazon ratings

A comparative study using Dictionaries approaches

by

Inês Bettencourt Martins Amorim

UNIVERSIDADE CATÓLICA PORTUGUESA

Using Sentiment Analysis to predict Amazon ratings

A comparative study using Dictionaries approaches

Final Work in Academic Context presented to

Universidade Católica Portuguesa

in order to obtain the master's degree in business economics

by

Inês Bettencourt Martins Amorim

Under the guidance of

Paulo Alves (PhD)

Católica Porto Business School, Universidade Católica do Porto

September, 2023

Resumo

Esta dissertação aborda o tema de *Sentiment Analysis*, uma técnica que permite detetar e extrair sentimentos humanos a partir de texto. Com o crescimento exponencial de dados sob a forma de texto online, particularmente nas avaliações dos consumidores, a necessidade de determinar com precisão os sentimentos destes nunca foi tão imperativo. Esta técnica é essencial para converter os dados textuais em informação que pode ser efetivamente utilizada. Para explorar a eficácia dos modelos de *Sentiment Analysis* na categoria de abordagem por Dicionário, este estudo implementa nove modelos: VADER, TextBlob, NRC Lexicon, SentiWordNet, Pattern, AFINN, Opinion Lexicon, LabMT e ANEW. Estes modelos são testados numa base de dados que contém avaliações da Amazon e classificações através das quais a precisão da extração de sentimento pode ser avaliada. O estudo aprofunda-se numa análise comparativa, avaliando o desempenho destes modelos para identificar os seus pontos fortes, fracos e a sua utilidade.

Palavras-Chave: Sentimento, Dicionário, Avaliações, Classificações, Amazon, Comparação

Abstract

This dissertation delves into the domain of sentiment analysis, a computational approach to detect and extract human sentiments from textual data. With the ever-increasing growth of online textual content, especially in the form of reviews, the need to accurately determine customer sentiment has never been more imperative. To explore the efficacy of lexicon-based sentiment analysis models, this study implements 9 models: VADER, TextBlob, NRC Lexicon, SentiWordNet, Pattern, AFINN, Opinion Lexicon, LabMT, and ANEW. These models are tested on an Amazon reviews dataset, which is uniquely accompanied by a rating system in which the accuracy of the sentiment extraction can be assessed. The study then further delves into a comparative analysis, collecting the performance of these models to discern their strengths, weaknesses, and overall utility.

Keywords: Sentiment Analysis, Dictionary approach, Reviews, Ratings, Amazon, Comparison

Acknowledgements

I would like to thank my parents for their unwavering support throughout my academic journey and for giving me the life-changing opportunity to study abroad, which had been a dream of mine for years. I am also deeply thankful to my little brother, my friends and my boyfriend, whose emotional support has been invaluable during this time.

I would like to express immense gratitude to my supervisor Paulo Alves, for his insights and guidance throughout this project.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

- ANEW: Affective Norms for English Words
- API: Application Programming Interface
- CNN: Convolutional Neural Network
- HITs: Human Intelligence Tasks
- KWIC: Keyword in Context
- LabMT: Language Assessment by Mechanical Turk
- MTurk: Amazon Mechanical Turk
- NLTK: Natural Language Tool Kit
- NLP: Natural Language Processing
- NRC: National Research Council
- POS: Part-of-Speech
- RNN: Recurrent Neural Network
- SVM: Support Vector Machine
- Synsets: sets of synonyms
- TF-IDF: Term Frequency- Inverse Document Frequency
- VADER: Valence Aware Dictionary and Sentiment Reasoner

Chapter 1: Introduction

The explosive growth of digital information has led to an unprecedented abundance of data (Parashar et al., 2013). Notably, a substantial fraction of this data is unstructured, often in the form of text derived from diverse origins such as social media, customer evaluations, and online discussion platforms in which people increasingly use technology to understand others' perspectives (Prager, 2006). This unstructured information has significant underlying potential for companies to increase customer experience and overall business performance, but its nature, mostly text, also imposes significant challenges and requires new analysis techniques and tools (Parashar et al., 2013).

Recognizing these challenges, several innovative techniques and tools have been developed in recent years to analyse textual data. These methodologies aim to understand people's sentiments, opinions, and intentions and uncover hidden patterns, trends, and themes. Among them, sentiment analysis emerges as a powerful technique that interprets and classifies emotions embedded within textual data. This technique's mounting interest and wide-ranging applicability have fostered a corresponding surge in the development of new methodologies in recent years (Ribeiro et al., 2016).

This dissertation aims to explore the realm of sentiment analysis by comparing different dictionary approaches. The chosen dataset contains reviews and ratings that represent the customer's sentiment. We plan to compare the efficacy of the different models by their ability to predict the ratings. The primary objective of this research is to augment the existing body of knowledge on sentiment analysis and derive insights that could pave the way for enhanced methodologies.

Online reviews and ratings are the backbone of e-commerce, shaping consumer perceptions and driving purchase decisions. Consequently, businesses are increasingly recognizing their potential and looking for ways to harness insights from customer reviews which makes its analysis paramount in today's digital age. Additionally, the choice of Amazon as the focal point of the project was due to its vast and diverse user base and dual mechanism of reviews and ratings.

The adopted methods involve attributing a predetermined polarity score to each word in the text based on a predefined sentiment lexicon or dictionary. These scores reflect the sentiment value associated with each specific word. To determine the overall sentiment

of an entire review, the scores of all individual words are aggregated and then categorized as either positive, neutral, or negative.

The structure of the project is as follows - ‘Chapter 1’ introduces the topic, in 'Chapter 2: Literature Review', discusses the evolution of text analysis, defines Natural Language Processing and then focuses on Sentiment Analysis elucidating its levels, categories, and varied methodologies. It further explores performance measures, review analysis and the Amazon dataset. ‘Chapter 3: Research Methodology describes the specific models employed, detailing each one and elaborating on data sourcing, cleaning, and preliminary analysis. 'Chapter 4: Results and Data Analysis' presents a comprehensive breakdown of the results and comparing the outcomes of the different models. Finally, 'Chapter 5: Conclusion' offers recommendations and speculates on possible future research.

The findings of this research strongly affirm the merits of dictionary approaches in sentiment analysis. Notably, most of the methods chosen have proven to be effective and reliable, achieving high levels of prediction accuracy. One of their standout features is the ability to be rapidly deployed, allowing for swift insights into vast amounts of data. This contrasts with machine-learning techniques, which often necessitate intricate adjustments and significant computational power. So, even though, Dictionary Approaches face several challenges such as domain dependency and missing contextual nuances, the combination of accuracy and speed positions dictionary approaches in a privileged position for real-time analysis for informed decision-making processes.

Chapter 2: Literature Review

2.1 History of Text Analysis

The field of sentiment analysis, also known as opinion mining, finds its roots in the late 1990s and early 2000s and has since been highly researched and developed (Mäntylä et al., 2018). In its nascent stages, sentiment analysis was primarily used to analyse long texts, letters, and emails and post-crime analysis of criminal activities using basic techniques such as keyword spotting and manual coding (Patel, 2015). Over the years, the research started focusing on different domains such as product reviews, predicting financial markets and analysing reactions to terrorist attacks and tackled more complex issues like detecting irony and supporting multiple languages (Mäntylä et al., 2018).

Natural Language Processing (NLP) played a pivotal role in enhancing Sentiment Analysis' accuracy and scope. Together, they have reshaped the landscape of textual data analysis over the past decades (Pang et al., 2002).

2.2 Natural Language Processing

NLP is a widely used term and it refers to the set of techniques and algorithms used to computationally analyse and understand human language in textual form. NLP techniques such as Tokenization, word segmentation, Part of Speech (POS) tagging, and parsing are examples of pre-processing steps to structure the text and extract features (Sun et al., 2017).

NLP techniques offer transformative advantages, including the rapid and consistent analysis of vast amounts of textual data by deciphering and learning from the abundant human language content online. They can also facilitate human-to-human exchanges via machine translation and boost human-machine interactions through conversational agents (Hirschberg & Manning, 2015). However, NLP is also challenged by inherent ambiguities in human language, nuances of sarcasm, cultural references, and the continuous evolution of language. Initially, most of the existing techniques only work for the English language but we are now observing a fast development in other languages (Sun et al., 2017).

Examples of NLP in action include speech recognition systems, such as Google Voice Search, content comprehension algorithms that can summarize or categorize text, and sentiment analysis tools that interpret the emotions expressed in text[1].

One of the many applications of NLP is in sentiment analysis, the object of this dissertation.

2.3 Sentiment Analysis

The explosive growth of social media platforms has prompted individuals, as consumers, and organizations, as manufacturers, to increasingly rely on public opinion for their decision-making processes. Paired with the intense competition in today's global market, the development of sophisticated mechanisms such as sentiment analysis has become crucial to capitalize on customer feedback (Patel, 2015).

Until recently, the majority of text analysis research was primarily geared toward extracting factual information through methods such as text classification and clustering. However, the current need is to understand subjective opinions, a task that proves to be more complex due to its inherently subjective nature (Liu, 2010). Sentiment analysis has become a go-to tool for marketers, providing a computer-assisted, fast, scalable, and efficient method of assessing consumer sentiment (Dhaoui et al., 2017).

Sentiment analysis's primary objective is to discern subjective information from textual data - determining the sentiment expressed in a text. This can also extend to measuring the intensity of the sentiment expressed (Taboada et al., 2011).

Sentiment classification falls into two main categories: subjectivity analysis and polarity analysis. Subjectivity analysis aims to distinguish between subjective and objective expressions in the text. Subjective expressions are those that convey personal opinions, emotions, or judgments, while objective expressions present factual information without

[1] To illustrate the capabilities of NLP, practitioners often employ libraries like the Natural Language Toolkit (NLTK). NLTK is one of the most prominent libraries for handling human language data. It provides a suite of tools and Python packages for a wide range of tasks in NLP, including classification, tokenization, and lemmatization of text. Additionally, NLTK offers user-friendly interfaces to complex linguistic databases and resources, such as WordNet and TextBlob (Abiola et al., 2023).

showing personal feelings or opinions. On the other hand, polarity analysis focuses on categorizing the sentiment expressed in the text as positive, negative, or neutral. Positive sentiment indicates a favourable opinion or feeling, negative sentiment suggests an unfavourable opinion or feeling, and neutral sentiment implies the absence of strong emotions or opinions in either direction. This classification helps in understanding the tone and emotional stance of a given piece of text (Koto & Adriani, 2015).

2.3.1 Levels of Sentiment Analysis

Sentiment analysis can operate at various levels (document, sentence, and aspect/feature) depending on the complexity and needs of the task at hand. The document-level analysis classifies the overall polarity of an entire text, often used when a single entity is being discussed by one author. The sentence-level analysis assigns sentiments to individual sentences within a document, beneficial for texts discussing multiple subjects. The aspect or feature level identifies sentiments tied to specific elements within the text (Fang & Zhan, 2015; Hardeniya & Borikar, 2016)[2].

2.3.2 Categories of Sentiment Analysis

Sentiment analysis can be classified into two main categories: machine learning - the supervised learning approach, and the lexical-based method, also known as the dictionary approach. There is also the hybrid method that combines the two previous methods (Hartmann et al., 2023).

2.3.2.1 Machine Learning Method

The machine learning approach involves several stages: data collection, pre-processing, training data, classification, and results plotting. The sentiment classifier is trained on a set of data, creating a model that is subsequently used for classifying new or unseen text.

[2] Although less common, there are also other levels such as the parts of speech and the weighted scheme level (Ashok Kumar & Abirami, 2015).

The accuracy of this classifier is greatly determined by the careful selection of features, typically unigrams, bigrams, or trigrams (Patel, 2015).

Feature extraction often involves converting raw text into numerical or categorical values to feed the algorithm. This conversion process might involve a variety of techniques, such as Bag of Words, Term Frequency-Inverse Document Frequency (TF-IDF), n-grams, or even more complex methods like word embeddings (Shrestha & Nasoz, 2019).

This method utilizes specific algorithms such as Support Vector Machines (SVM) and Naïve Bayes and maximum entropy classification (Pang et al., 2002). There has been also a rise in deep learning techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which are increasingly used in sentiment analysis due to their ability to capture complex patterns and understand the context more effectively (Sun et al., 2017).

Machine learning is increasingly utilized in sentiment analysis due to its adaptability and precision. However, this method often requires considerable resources and labelled data for training. High-quality labelling of a vast training dataset can be a time-consuming process, and reducing the dataset size can negatively affect classification accuracy. (Dhaoui et al., 2017).

2.3.2.2 Lexicon Method

Dictionary approaches offer a simpler, faster, and less resource-intensive alternative. The process starts with preprocessing, which prepares the text for analysis through tokenization, stemming, and stop-word removal which translates into n-gram representation. A sentiment-carrying lexicon or dictionary, either pre-existing or custom-built, is then utilized. Each text is examined, with sentiment scores from the lexicon combined to give an overall sentiment score. The text is subsequently classified as positive, negative, or neutral based on this score (Hardeniya & Borikar, 2016).

Although simple to implement and intuitive, this approach faces several challenges. First, a word's sentiment orientation can flip depending on the context or domain, making a fixed sentiment assignment tricky (e.g., liability has often a negative association, but in a financial realm is a technical term). Additionally, not all words bear sentiment, and the lexicon-based approach may incorrectly interpret these neutral words, leading to skewed

sentiment scores. Furthermore, this method predominantly focuses on subjective sentiments, making it difficult to identify and extract objective information from the text (Hardeniya & Borikar, 2016). Other issues include struggling with sarcasm or irony, dealing with fixed expressions composed of multiple words that convey sentiments not equal to the sum of individual word sentiments, and keeping up with the constant evolution of language, including new slang and abbreviations (Pedrycz et al., 2016).

Taboada et al. (2011) found that dictionary approaches can achieve similar accuracy to machine learning methods for sentiment analysis in movie reviews. Kiritchenko et al., (2014) created tweet-specific lexicons that outperform in sentiment prediction on tweets, under both unsupervised and supervised conditions. This highlights the need for further research to understand the contexts in which dictionary-based methods are most effective.

Examples of some of the most popular tools include VADER, TextBlob, NRC emotion lexicon, SentiWordNet, Pattern, AFINN, Opinion Lexicon, LabMT, and ANEW. This dissertation will shed light on the utility of these models by testing their ability to accurately predict customer ratings in the selected Amazon dataset.

2.3.2.3 Hybrid Method

The hybrid method of sentiment analysis merges the machine learning and dictionary methods, aiming to capitalize on the strengths of both while mitigating their respective weaknesses (Patel, 2015). For instance, it might use a dictionary approach as a baseline or as a feature within a more comprehensive machine learning model. This combination can result in a system that is both fast and highly accurate. Numerous studies such as Dhaoui et al. (2017) have demonstrated that this hybrid approach can significantly improve the performance of sentiment classification tasks, making it a compelling choice for researchers and practitioners.

2.4 Performance Measures

To evaluate the performance of sentiment analysis methods, researchers commonly employ a set of metrics, including precision, recall, F1-score, and accuracy. Precision is

defined as the ratio of correctly predicted positive sentiments to the total instances predicted as positive. It essentially measures how many of the items identified as positive are positive. Recall calculates the proportion of actual positive cases that the model correctly identified, focusing on the model's ability to capture relevant instances. The F1-score, or F-Measure, serves as a single metric that combines both precision and recall through their harmonic mean, offering a balance between precision and recall. It ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 means neither precision nor recall. Lastly, accuracy calculates the proportion of all instances that the model correctly identified (Aljuhani & Alghamdi, 2019)[3].

When evaluating sentiment analysis models, two additional metrics that researchers frequently consider are coverage and agreement. Coverage refers to the proportion of text for which the model can generate a sentiment label. In dictionary approaches, this may involve assessing the percentage of words in the input text present in the sentiment lexicon. High coverage is desirable, as it signifies that the model can provide sentiment scores for a substantial portion of the input data. Agreement is a metric that quantifies the alignment between the sentiments identified by the model and the actual sentiments, or the ground truth. High agreement indicates that the model's predictions are not only frequent but also accurate and trustworthy, reflecting the model's capability to correctly interpret the emotional tone of the text (Gonçalves et al., 2013).

2.5 Review Analysis

In today's interconnected world, reviews have become immensely important, not only due to the explosion of online shopping but also because people rely on these reviews for all kinds of purchases. These reviews, which are plentiful as most businesses enable users to share their opinions, represent a rich source of data. They offer invaluable insights into customer satisfaction, product quality, and service experiences, among other aspects. As the volume of these reviews continues to grow exponentially, so does the necessity for effective tools and methodologies to analyse and extract meaningful, actionable information. Businesses are recognizing the potential of harnessing this wealth of

[3] While accuracy can provide a useful general measure, it can be misleading in imbalanced datasets since it does not account for false positives and false negatives as explicitly as precision, recall, and F1 scores do.

consumer input and therefore increasingly seeking solutions that would allow them to gain valuable insights from customer reviews. These insights be valuable for product development and marketing strategies, and overall customer engagement efforts, thereby allowing companies to better meet the needs and expectations of their customers (Aljuhani & Alghamdi, 2019; Tanjim Ul Haque et al., 2018).

Reviews, being rich with subjective opinions and sentiments, present an excellent opportunity for the development and validation of sentiment analysis techniques. These datasets allow researchers to rigorously test their models under real-world conditions, thereby ensuring the scalability and robustness of their approaches. Ohana et al. (2009), for example, employed the dictionary approach for the analysis of film reviews using SentiWordNet. Achieving an accuracy of 65.85%, their findings suggest that this lexicon not only performs well compared to manual resources but also enhances the efficacy of supervised models when integrated. Ultimately, their research underscores the utility and effectiveness of the lexicon in sentiment analysis. In contrast, Baid et al. (2017), adopted machine learning methods for movie reviews. Their accuracies varied, with K-Nearest Neighbour yielding 55% and Random Forest achieving 81.4%. Their primary takeaways highlight the relevance of determining review polarity and suggest exploring hybrid methodologies in future endeavours to enhance accuracy. Nonetheless, a common thread between these studies and many others is the recognized utility of lexicons in sentiment analysis and the relevance of review analysis.

2.6 Amazon Dataset

Amazon's global reach and the dual mechanism of textual reviews paired with ratings provide a robust framework for validating sentiment extraction methodologies. The textual reviews offer qualitative insights into customer feelings, experiences, and concerns, while the ratings provide a quantitative measure that can be directly associated with these sentiments (Shrestha & Nasoz, 2019).

Several studies have capitalized on this rich dataset that Amazon provides. Hamouda & Rohaim (2011) recorded an accuracy of 56.77%, suggesting that adopting certain techniques, like omitting words with high neutral scores, could enhance model accuracy and reinforce the utility of the lexicon in sentiment analysis.

Shrestha & Nasoz (2019) research harnessed the power of machine learning methods and advanced deep learning techniques, using paragraph vectors, RNN, Gated Recurrent Unit, and then SVM. The objective of their study was to demonstrate that the incorporation of product embeddings could enhance prediction accuracy, which was validated with an improvement in accuracy from 81.23% to 82.82%.

Most of the studies conducted on Amazon reviews have predominantly employed machine learning methods, with some incorporating semi-supervised techniques that utilize lexicon tools as supplementary resources. Notably, there has been limited research focused solely on using the lexicon approach for sentiment analysis within this context.

Chapter 3: Research Methodology

In this dissertation, similarly to the work of Gonçalves et al. (2013) on tweets, we adopted the Dictionary approach, and we'll be using Python as the primary programming language.

3.1 Models

The objective is to assess and compare the performance of the chosen models: VADER, TextBlob, NRC emotion lexicon, SentiWordNet, Pattern, AFINN, Opinion Lexicon, LabMT, and ANEW.

3.1.1 VADER

The Valence Aware Dictionary and Sentiment Reasoner (VADER) sentiment analysis tool was introduced by Hutto and Gilbert (2014). Designed to decipher sentiment in short, informal text, VADER is particularly effective for analysing modern forms of communication that are replete with slang, emoticons, and acronyms. This makes it exceptionally attuned to microblog-like contexts, such as Twitter and other social media platforms. To construct VADER, Hutto and Gilbert (2014) employed a meticulously crafted lexicon and incorporated five generalizable rules that embody the grammatical and syntactical conventions humans use when expressing sentiment intensity. These rules enable the model to account for intensifiers, negations, and other linguistic elements that influence the sentiment of a phrase. VADER provides a fine-grained analysis of text sentiment, effectively distinguishing between different levels of positivity or negativity. It has proven to be a powerful tool in the field of sentiment analysis (Hutto & Gilbert, 2014).

VADER has garnered significant attention in the research community and has even been implemented as part of the NLTK Python library (Ribeiro et al., 2016). Numerous studies have validated its reliability in sentiment analysis tasks across diverse domains.

For example, Ribeiro et al. (2016) evaluated various sentiment analysis methods, including VADER. According to their findings, VADER's accuracy varied significantly depending on the context and dataset (49% to 99%). The study also noted instances where VADER's coverage fell below 30%, indicating that the model could not confidently analyse a substantial portion of the given texts. Despite this, the paper confirmed VADER's strength in analysing sentiment in social media contexts.

3.1.2 TextBlob

The TextBlob library is a package for Python that provides a consistent API (Application Programming Interface) for diving into common NLP tasks. It employs a sentiment dictionary alongside the pattern.en sentiment evaluation mechanism. Pattern.en[4] uses WordNet to determine sentiment based on the English adjectives present in the content. When TextBlob conducts a sentiment assessment on a piece of text, it provides a pair of values in the format (polarity, subjectivity), with polarity being a number between [-1,1] (Sohangir et al., 2018). This technique became a popular tool among researchers and developers due to its ease of use and extensive functionalities, which include, but are not limited to, sentiment analysis. Other applications such as part-of-speech tagging, noun phrase extraction, classification and translation (Abiola et al., 2023).

TextBlob offers a range of advantages for sentiment analysis but also poses several notable challenges. The library only provides measures of polarity and subjectivity, which may not be sufficient for all analytical purposes. It may have difficulty interpreting emojis and might not be the ideal tool for analysing biased or skewed reviews, as it may struggle to detect nuanced or subtle sentiments. Additionally, TextBlob may face complications when analysing text in multiple languages, which could limit its effectiveness in assessing emotions across diverse linguistic contexts (Praveen Gujjar & Prasanna Kumar, 2021).

In a sentiment analysis study conducted by Sri et al. (2019), TextBlob was compared to VADER to assess sentiments in patient feedback in healthcare. TextBlob marginally outperformed VADER in accuracy (73% to 71.9%), but a comprehensive evaluation of other metrics led the researchers to conclude that VADER performed the best. Still, both

[4] The details of the Pattern Library will be explained in a later section, in the chapter discussing the Pattern model.

models demonstrated their effectiveness in extracting patient sentiments in the context of this research.

3.1.3 NRC Emotion Lexicon

To create The NRC Emotion Lexicon, also known as EmoLex, Mohammad and Turney employed the Amazon Mechanical Turk (MTurk) platform, a marketplace where 'human intelligence tasks' (HITs) are performed by workers referred to as ‘requesters’ for compensation. These tasks, which can range from image annotation to speech transcription, are typically simple for humans but challenging for computers. Using this platform, the authors designed specific HITs to help them build a lexicon that can identify 10 distinct emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive and negative.[5] (Akkaya et al., 2010; Mohammad & Turney, 2010).

The NRC Emotion Lexicon has since become one of the most used lexicons in sentiment analysis, but it poses a series of issues. For example, according to Zad et al. (2021), the lexicon does not specify the part of speech for terms associated with emotions, leading to ambiguity in understanding whether a specific emotion label is relevant to a particular use of a word form. It also contains various incorrect or nonsensical associations, often resulting from a conflation of emotive and affective emotional language, thereby neglecting the role of context and it includes emotion markings that lack supporting evidence in Keyword-in-Context (KWIC) databases, indicating potential errors in the lexicon.

3.1.4 SentiWordNet

SentiWordNet was introduced by Sebastiani & Esuli (2016) as an extension of WordNet, a previous model developed by the authors that maps English words into sets of synonyms called synsets. SentiWordNet aims to provide sentiment scores for each WordNet synset and assign to each synset three sentiment scores: positivity, negativity, and objectivity. The creation of SentiWordNet involved the use of machine learning techniques to

[5] An additional class, named ‘No emotion’, is created when none of the 10 emotions is detected.

automatically assign sentiment scores to the synsets based on the features derived from the glosses (dictionary definitions and examples) associated with each synset. According to the authors, one of the most significant advantages of this model is its extensive coverage, which stems from the integration of WordNet synsets.

The SentiWordNet lexicon has established itself as a valuable resource for opinion mining, with its applications extending to various research domains (Ohana et al., 2009). For example, Hamouda & Rohaim (2011) used SentiWordNet in sentiment classification of film reviews and found that its accuracy ranged from 60% to 68% and that performs well compared to manual lexicons. Additionally, using this model as a feature source in a supervised learning model demonstrated improvements over the basic word-counting approach.

3.1.5 Pattern

The Pattern Library is a versatile web mining package for Python that offers tools for NLP, Machine Learning, Network Analysis, graph visualization and more. Pattern's sentiment analysis module uses a lexicon-based approach, leveraging a built-in lexicon of adjectives that are frequently used to express sentiment. Each adjective is assigned a sentiment polarity score and a subjectivity score. The sentiment function in Pattern returns a (polarity, subjectivity) tuple for a given input text, where polarity is a float in the range of -1 (negative) to 1 (positive), and subjectivity is a float in the range of 0 (objective) to 1 (subjective) (De Smedt & Daelemans, 2012).

Because of its versatility and ease of use, Pattern has become a popular choice among researchers and developers for various tasks, including sentiment analysis, part-of-speech tagging, and text classification, among other NLP tasks. As an illustrative example, Filho & Pardo (2013) employed the Pattern library as a part-of-speech tagger in their hybrid system, which was designed to perform sentiment analysis on Twitter messages.

3.1.6 AFINN

The AFINN model, introduced by Finn Årup Nielsen in 2011, utilizes a precompiled list of English words, which was originally created in 2009 for tweets, and each word is assigned an integer score indicating its sentiment polarity ranging from -5 (negative) to +5 (positive). The scores of all the words in the text are then summed to obtain an overall sentiment score for the text. A positive sum indicates a positive sentiment, a negative sum indicates a negative sentiment, and a sum close to zero indicates a neutral sentiment. The model's simplicity and efficiency have contributed to its widespread adoption, especially in analysing social media content where short and informal expressions prevail. One of the model's notable strengths is its inclusion of internet slang and obscene words, enhancing its sensitivity to the nuances of contemporary, online discourse. However, the AFINN model has notable limitations: it struggles to accurately detect negations, does not interpret emoticons and is not sensitive to spelling variations (Nielsen, 2011).

Subsequent research, which also focused on analysing Twitter content, has confirmed that the AFINN model performs commendably in sentiment analysis tasks (Koto & Adriani, 2015).

3.1.7 Opinion Lexicon

Opinion Lexicon, also known as Bing Liu's Lexicon, is part of the NLTK library and is a popular resource for sentiment analysis tasks. It was created by Minqing Hu and Bing Liu and they used WordNet and Part-of-Speech Tagging (POS) to create a seed list of positive and negative words and then updated it according to the domain they were analysing – customer reviews (Hu & Liu, 2004).

The model is essentially two lists of words that are pre-classified as either positive or negative, and each contain words that typically express a positive sentiment and negative sentiment, respectively. When using this lexicon, a given input text is scanned for the presence of words from both lists. Each occurrence of a word from the positive list is counted as a positive sentiment, while each word from the negative list is counted as a negative sentiment. The overall sentiment of the text is subsequently determined by comparing these counts. Notably, the Opinion Lexicon does not assign numerical scores

to words instead, it simply classifies words as positive or negative, which makes it a straightforward and easy-to-use resource (Hu & Liu, 2004).

The Opinion Lexicon, while a valuable resource, has notable limitations. It offers only binary classification of words as positive or negative, without quantifying sentiment strength, and it does not account for the context in which a word is used, potentially leading to misinterpretations. Additionally, the lexicon may struggle with handling negations, and it may become outdated if not regularly updated to include new sentiment-bearing words. The lexicon also lacks categories for neutral or objective expressions and does not account for emoticons and emojis, which are significant in modern online communication (Saif M, 2017).

3.1.8 LabMT

The Language Assessment by Mechanical Turk (LabMT), also known as LabMT, is based on the assessment of word happiness by human evaluators using MTurk service. The lexicon contains over 10,000 words commonly used in the English language, which were sourced from Twitter, Google Books, music lyrics and the New York Times. Each word in the LabMT lexicon is associated with a happiness score that indicates the perceived happiness level of the word, as judged by human evaluators on a scale from 1 (least happy) to 9 (most happy). The average of these scores is then computed, yielding a single happiness score for the entire text (Mitchell et al., 2013).

The LabMT sentiment analysis approach, while effective for analysing large datasets, has several limitations. For small texts, especially at the sentence level, it may struggle to accurately judge tone or meaning due to inherent ambiguities. The technique extracts a perceived happiness level based on word frequency alone, intentionally simplifying the analysis compared to more complex NLP algorithms, and thus misses the structural and contextual nuances of the text. Furthermore, LabMT measures only the apparent emotional tone exhibited in written communication; it does not claim to access or understand the internal emotional states of individuals or populations(Dodds et al., 2011).

3.1.9 ANEW

The Affective Norms for English Words, more commonly known as ANEW was conceived to establish a standardized set of emotional ratings for a substantial collection of words in the English language (Bradley & Lang, 1999).

In terms of its structure, the ANEW lexicon consists of a collection of words, each of which has been rated by human subjects along three distinct emotional dimensions: Valence (ranging from pleasant to unpleasant), Arousal (ranging from calm to excited), and Dominance (degree of control). The ratings from the participants are then averaged to yield a normative score for each word (Andrea et al., 2015).

The ANEW model serves a multifaceted purpose, enabling researchers across various domains—including psychology, linguistics, and marketing—to assess the emotional tone of different types of text. By calculating the average valence, arousal, and dominance scores of words in a text, the model provides an estimate of the text's overall emotional content along these three dimensions. The ANEW model's scoring has been validated across several participants, reinforcing its credibility and widespread use in emotion research. It is part of a larger set of resources developed to study emotional responses to language, and its meticulous data collection and straightforward approach have established it as a prominent tool in the field (Nielsen, 2011).

3.2 Data

The dataset for this project was sourced from Kaggle (Datafiniti, 2019), and provides a comprehensive overview of product reviews sourced from Amazon. It contains 34,660 observations and 21 attributes or features (Table 1).

Field Name	Description
id	NaN
name	The product's name
asins	Amazon identifier used for this product
brand	The brand name of this product
categories	A list of category keywords
keys	A list of internal Datafiniti identifiers
manufacturer	The manufacturer of this product
reviews.date	The date the review was posted.
reviews.dateAdded	The date this product was first added
reviews.dateSeen	The date when the review was seen
reviews.didpurchase	Whether the reviewer purchased the product
reviews.datedoRecommend	Whether the reviewer recommends the product
reviews.id	The website ID associated with this review
reviews.numHelpful	Number of people that found the review helpful
reviews.rating	A 1 to 5 start value for the review
reviews.sourceURLs	A list of URLs where this review was seen
reviews.text	The full (or available) text of the review
reviews.title	The review's title
reviews.userCity	The reviewer's city
reviews.userProvince	The reviewer's province or state
reviews.username	The reviewer's username

Table 1 – Feature Description

The focus of our analysis will be mainly around 2 features: ‘reviews.text’ - the text data we will analyse - and ‘reviews.rating’ - numbers ranging from 1 to 5 - which will serve as the ground truth for the sentiment embedded in the review. However, there are other features in the dataset that can provide context for our analysis.

The data frame provides product details like the product's name ('name') and its corresponding Amazon identifier ('asins')[6]. Relying on the unique count of 'asins', the reviews refer to 41 different products.

[6] Ideally, ‘name’ and ‘asins’ should share an equal count of unique values, however, while 'name' has 48 unique entries, 'asins' records 41. This discrepancy suggests that certain product names might link to

The are also 6 brands, all of which are subdivisions of Amazon - Amazon, Amazon Fire, Amazon Echo, Amazon Coco T, Amazon Fire TV, and Amazon Digital Services Inc. and 2 manufacturers, also related to Amazon: Amazon, Amazon Digital Services, Inc. There are 41 product categories in total. Figure 1 illustrates the top 10 categories based on review count, highlighting that most of the products reviewed are electronic devices.

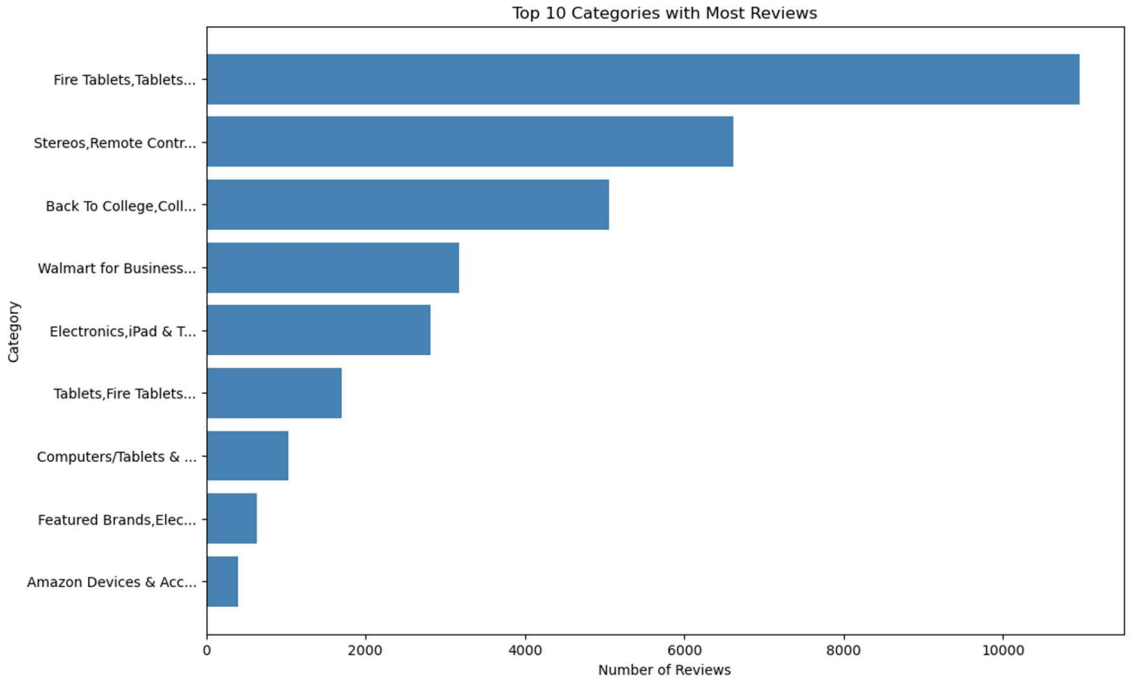

Figure 1 – Top 10 Categories with Most Reviews

The reviews themselves are enriched with metadata including information about the timeline they were posted ('reviews.date'), when they were added ('reviews.dateAdded'), and when they were last seen by Datafiniti ('reviews.dateSeen').

multiple identifiers, possibly due to the lengthiness of the names, increasing the likelihood of typographical errors or variations. Therefore, we will trust the unique number of 'asins' to assess the number of different products in the dataset.

From the ‘reviews.dateAdded’ variable, we were able to determine that the reviews were written between 2011-11-08 and 2018-04-18. Figures 2 and 3 represent the evolution in the number of reviews written and the average ratings over time, respectively. We can observe that most of the reviews were written between 2016 and 2018 and a predominant trend towards positive feedback.

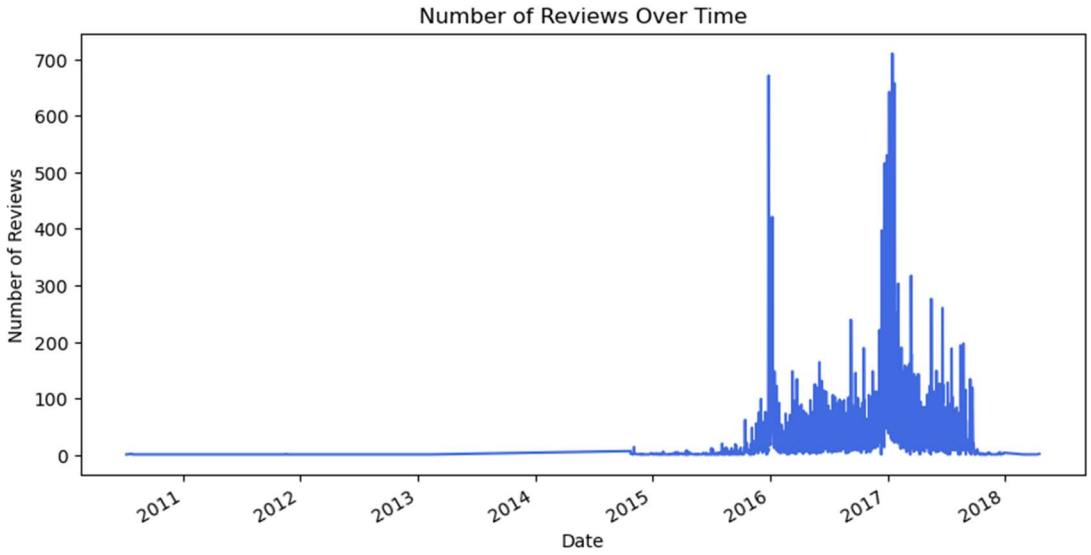

Figure 2 – Number of Reviews Over Time

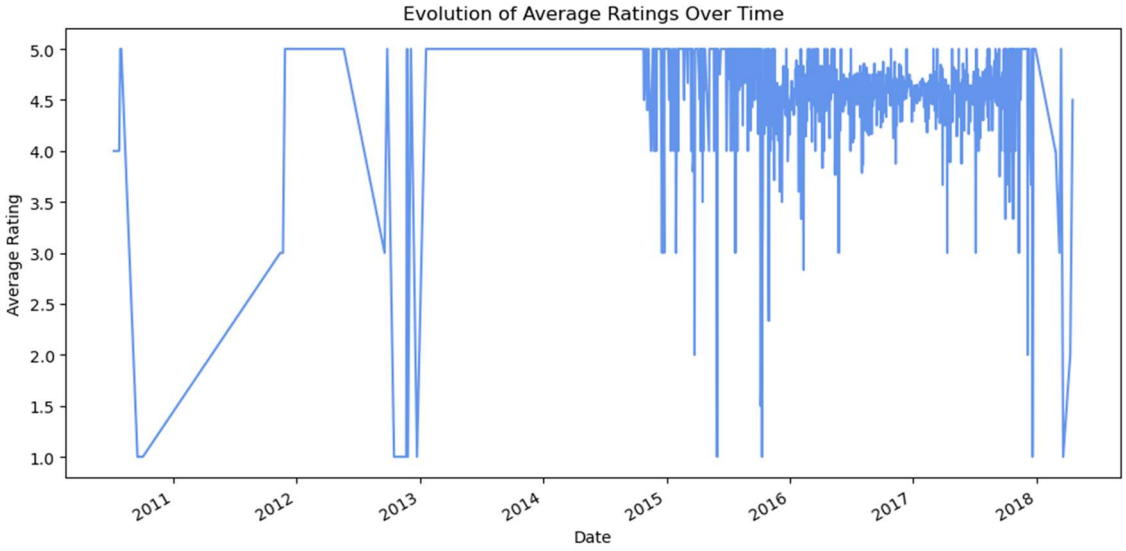

Figure 3 – Evolution of Average Ratings Over Time

We delved into the ‘reviews.doRecommend’ and ‘reviews.numHelpful’ attributes, finding that the majority of reviewers endorsed the product. Interestingly, as ratings decreased, the count of 'helpful' votes grew, suggesting that lower-rated reviews were deemed more helpful (Figure 4 and Figure 5).

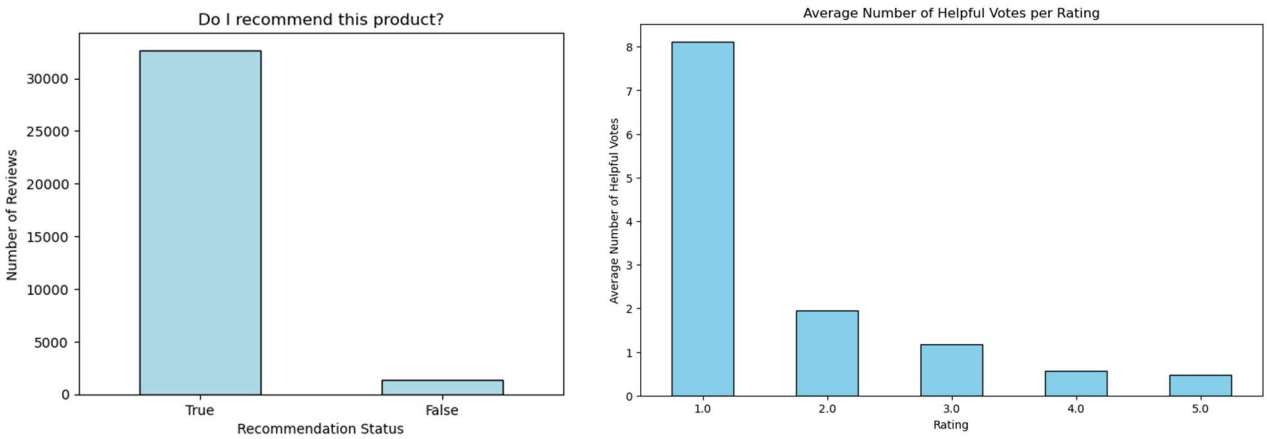

Figure 4 – Recommendations | Average Number of Helpful Votes per Rating

The dataset contains 26,788 distinct usernames, indicating that the reviews were written by that many unique individuals[7].

3.3 Data Cleaning and Pre-Processing

The only features that we will be working with from now on are the ‘review.text’ and ‘review.rating’, so all the remaining features were dropped from the dataset and renamed them ‘reviews’ and ‘ratings’, respectively, for simplicity.

To ensure quality and reliability, 33 observations were deleted due to missing data and duplicate records.

Furthermore, 'reviews' underwent thorough preprocessing to ensure is well prepared for model application later. Firstly, it's imperative to ensure that the text under examination consists solely of words, thereby filtering out any 'noisy' data. Therefore, HTML tags, content within square brackets, punctuation marks, special characters and numbers, which

[7] The remaining features do not add value to our analysis because they contain predominantly null values and/or are not relevant to our analytical goals.

might confound the models, were excised with regular expressions. Lastly, we converted all the text to lowercase to ensure uniformity.

Next, we used several NLP techniques to further prepare the data and make it amenable to computational methods. We embarked on the process of tokenization, which entails breaking down the text into individual words or tokens. It was also paramount to ensure the consistency of language across the dataset, so we implemented a language detection mechanism to retain only English-written content, maintaining the homogeneity of our data. And, to enhance the fidelity of our dataset, we incorporated a spell-checking tool that rectifies common misspellings, ensuring that each word is represented in its accurate form.

Lastly, to accentuate the significance of contextually relevant words, we employed a list of English stop words from the NLTK library. By eliminating these commonly used words, which often bear minimal semantic weight, we could distil the content of our reviews, spotlighting the terms most pertinent to sentiment and meaning.[8]

3.4 Exploratory Data Analysis

The following data analysis primarily aims to understand and visualize the content and distribution of product reviews. We started by identifying the cumulative word count of 492,822 words spanning all reviews. Concurrently, we spotlighted the vocabulary's distinctiveness, pinpointing 11,289 unique words.

In our analysis, we identified the most recurrent words in the reviews which are represented in the Word Cloud below (Figure 5). Notably, many of these terms carry a positive connotation, hinting at a largely favourable reception among reviewers.

[8] Additionally, we created a specialized stopword list that included the words: ‘tablet’, ‘amazon’, ‘alexa’, ‘kindle’, ‘books’ and ‘device’. These words were removed from the dataset because they appeared very frequently and would not contribute to the sentiment extraction process.

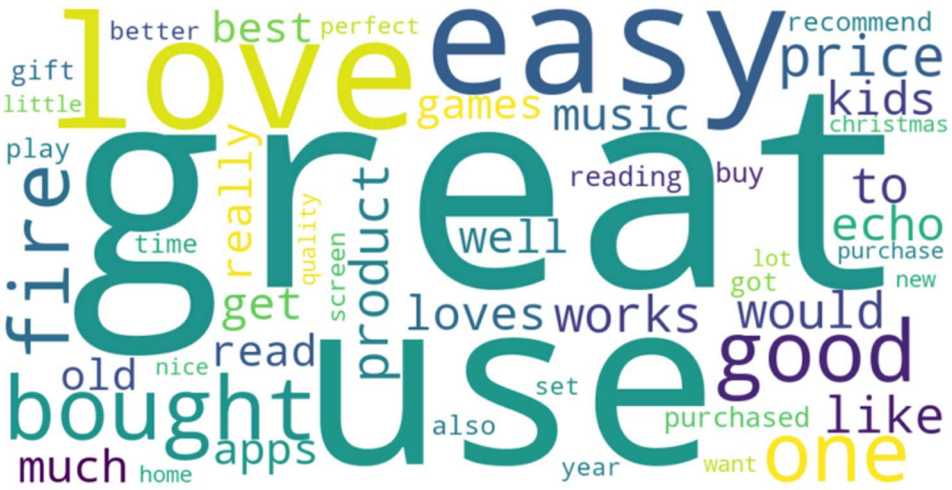

Figure 5 – WordCloud

Following the word-centric insights, we transition to the ratings by plotting the following graphs (Figure 6):

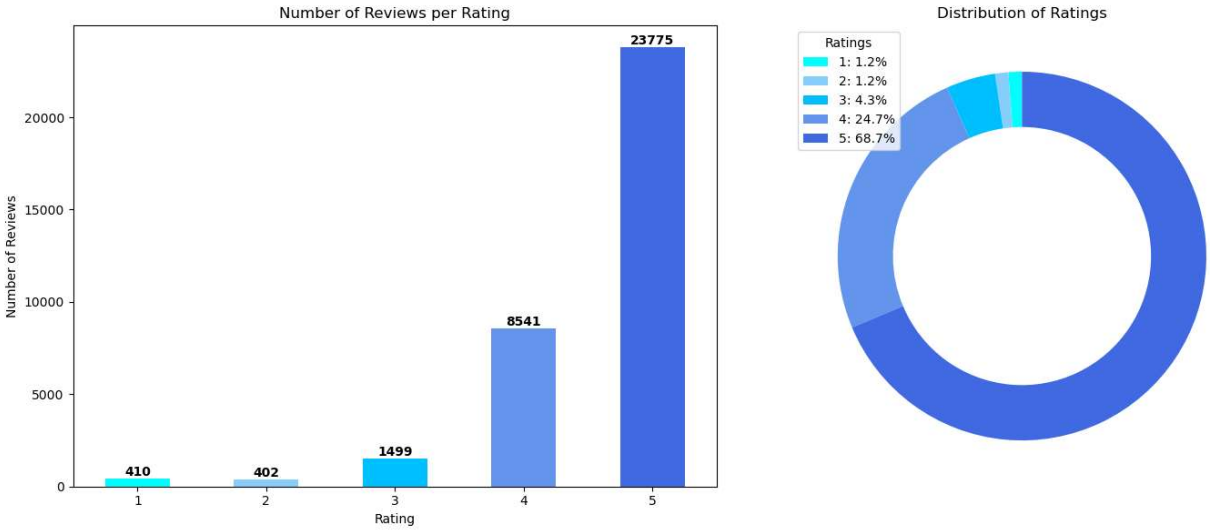

Figure 6 – Number of Reviews per Rating | Distribution of Ratings

Notably, the '5' rating dominates with 23,775 reviews, accounting for 68.7% of all reviews, followed by the '4' rating, which has been chosen 8,541 times, making up 24.7% of all reviews. This reinforces the conclusion that the vast majority of the reviews in our dataset lean towards the positive spectrum.

In essence, the analysis provides insights into the language used in product reviews and the distribution of customer ratings, setting the foundation for sentiment analysis.

Chapter 4: Results and Data Analysis

Our primary objective is to extract insights and offer a comprehensive view of the performance of the Lexicon-based models we selected. Having completed the data cleaning and pre-processing steps, we were able to implement each model and accumulate the results. This chapter focuses on the analysis and interpretation of these results, aiming to discern the advantages and disadvantages of the different sentiment analysis techniques.

4.2 Models – Results

Every model dissected the reviews' text and made predictions on the underlying sentiment. For a consistent comparison between the model's predictions and the given ratings, we standardized the outputs to the labels: 'Negative', 'Neutral', and 'Positive'. While each model had its own criteria for categorization, we interpreted ratings of 1 and 2 as 'Negative', 3 as 'Neutral', and 4 and 5 as 'Positive' sentiments for the entirety of our analysis.

A diverse range of metrics was calculated– accuracy, precision, recall, f1-score, – to assess their performance in predicting each class. However, to calculate the overall metric in datasets where certain classes are over-represented, as is the case here with more positive sentiments, a simple average might not truly capture the model's performance across all categories. By using a weighted average, each metric considers the proportion of instances from each class, providing a more balanced and realistic overview.

Overall, most of our selected models demonstrated commendable accuracy in their predictions. Additionally, we compared the performance metrics across models, calculated the level of agreement between them and unpacked their similarities and differences.

4.2.1 VADER

VADER attributed a polarity score ranging from -1 (negative) to 1 (positive) which was then used to assign each review to one class, as in Pano & Kashef (2020) : below -0.05 as 'Negative', above 0.05 as 'Positive', in between as 'Neutral'.

Figure 7 represents the distribution of sentiments attributed by VADER and the distribution of grouped ratings. We can observe that they present similar distributions, implying a high accuracy of the model's predictions. Notably, in both graphs, the 'positive' sentiment emerged as the predominant classification. The model identified 31,186 reviews as positive, accounting for 90.1% of the total, whereas in reality this sentiment represented 93.3%.

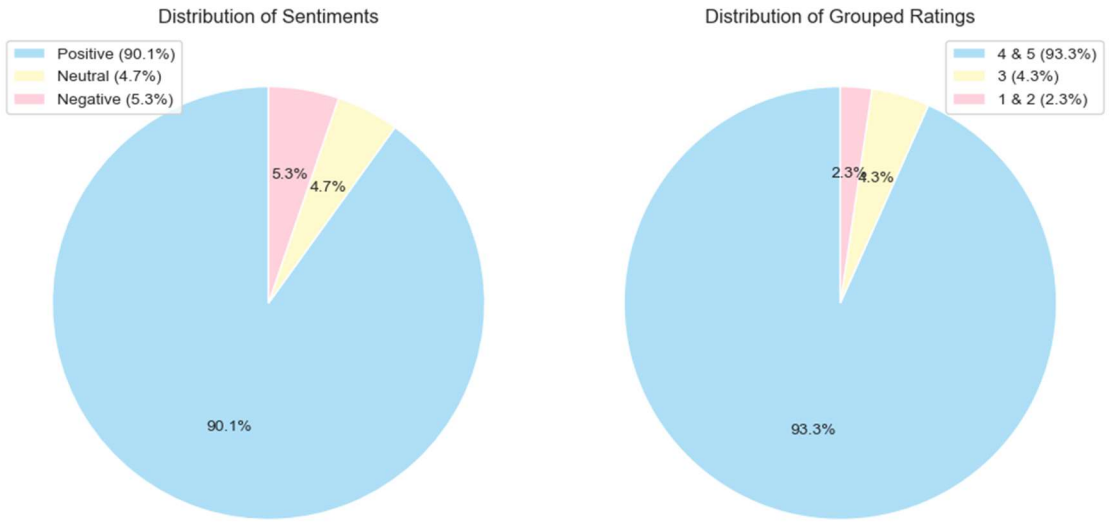

Figure 7 – VADER Distribution of Sentiments and Distribution of Grouped Ratings

We then determined whether the sentiment assigned by the model to each review correctly aligns with the given rating and got an accuracy of 86.55%.

We also calculated the precision, recall, and f1-score for each class (Table 2)[9]:

[9] The support column on the metrics report represents the percentage of actual occurrences.

	precision	recall	f1-score	support
Negative	14.01%	31.4%	19.38%	2.34%
Neutral	8.39%	9.07%	8.72%	4.33%
Positive	94.85%	91.53%	93.16%	93.33%
macro avg	39.08%	44.0%	40.42%	100.0%
weighted avg	89.21%	86.55%	87.77%	100.0%

Table 2 – VADER's Metrics Report

The report suggests that the model performs exceptionally well in identifying positive sentiments, achieving an F1-score of 93.16%. However, the model's performance for neutral sentiments is suboptimal. It has a low precision and recall, suggesting difficulty distinguishing neutral reviews. The performance on negative sentiments is also not very high. Although the recall is at 31.4%, indicating that it was able to identify a portion of the negative sentiments correctly, the precision is only 14.01% meaning that many reviews predicted as negative weren't negative.

Lastly, by calculating the model's coverage, we concluded that only about 13.28% of the vocabulary from the reviews is present in Vader's lexicon. This indicates potential limitations in Vader's ability to analyse and assign sentiments to certain words that exist in the reviews but are absent from its lexicon.

Overall, the results indicate an admirable performance: a precision of 89.21%, a recall of 86.55% and an F1-score of 87.77%.

These results are consistent with the literature, VADER was created to analyse short texts and informal language which justifies its good performance, and it often achieves a coverage below 30%.

4.2.2 TextBlob

The TextBlob model attributed a polarity score from -1 (Negative) and 1 (Positive) to each review which was transformed into sentiment labels using the following rule: scores above zero as 'Positive', below zero as 'Negative', and precisely at zero as 'Neutral'.

Figure 8 showcases the sentiment distribution determined by the model in comparison to the distribution of grouped ratings. Their resemblance suggests that the model's predictions align closely with the actual sentiments. Just as we noticed with the Vader model, the 'positive' sentiment is dominant in both distributions. The TextBlob model classified 30,840 reviews as positive, making up 89.1% of the total.

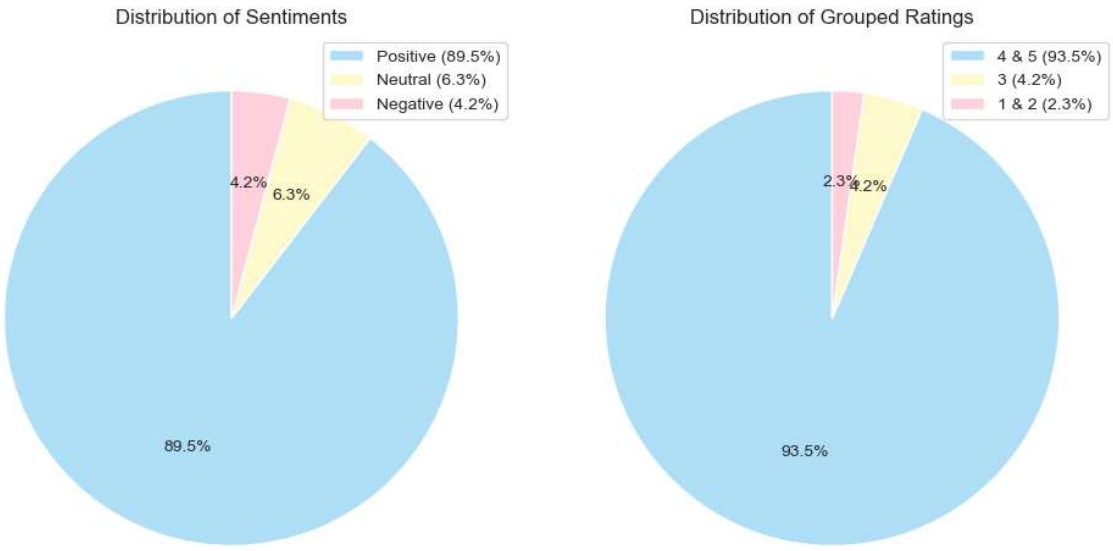

Figure 8 – TextBlob Distribution of Sentiments and Distribution of Grouped Ratings

By comparing the model's sentiment predictions with the actual ratings of each review, we calculated TextBlob's accuracy and concluded that it made accurate predictions for 85.13% of the reviews[10].

Regarding the performance metrics for each class, Table 3 reveals that the model is highly proficient at detecting positive sentiments, boasting a precision of 94.49% and a recall of 90.17% which results in an F1 score of 92.28%. However, when it comes to neutral and negative sentiments, the model's performance is notably lower. For neutral reviews, both precision and recall hover around the low values of 5% and 8% respectively. The model's

[10] In the case of this model, we couldn't compute the coverage due to our inability to access the model's lexicon.

ability to identify negative sentiments isn't very good either. Though the recall is 26.48%, suggesting a somewhat decent identification rate, the precision sits at a mere 14.88%.

	precision	recall	f1-score	support
Negative	14.88%	26.48%	19.05%	2.34%
Neutral	5.21%	8.14%	6.35%	4.33%
Positive	94.49%	90.17%	92.28%	93.33%
macro avg	38.19%	41.6%	39.23%	100.0%
weighted avg	88.76%	85.13%	86.84%	100.0%

Table 3 – TextBlob's Metrics Report

Overall, the TextBlob method's results showcase efficacy: a precision of 88.76%, a recall of 85.13% and an F1-score of 86.84%. It's worth noting that even though previous studies suggest that this model is not ideal for skewed data, it still achieved a harmonious performance across these key metrics.

4.1.3 NCR Emotion Lexicon

The NRC lexicon assigns a dominant emotion to each text entry offering a richer insight into the emotional tone. Figure 9 illustrates these results:

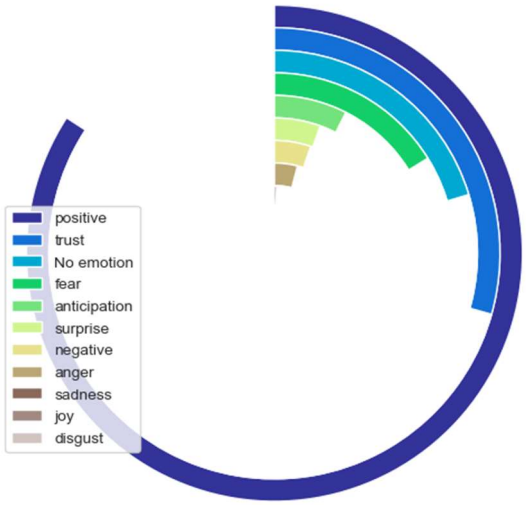

Figure 9 – NCR Sentiment Distribution

The emotions joy, trust, positive, anticipation, and surprise were grouped under the 'Positive' sentiment. Reviews where no discernible emotion and therefore labelled as 'No

emotion', were recategorized as 'Neutral'. All other emotions – disgust, anger, sadness, negative - which predominantly have negative connotations, were labelled as 'Negative'. This transformation permitted the comparison with the reviews' ratings illustrated in Figure 10. The NCR lexicon identified only 73.2% of the reviews as positive, in reality, 93.3% were positive. Furthermore, the lexicon misclassified a notable portion of reviews, wrongly tagging many as negative or neutral.

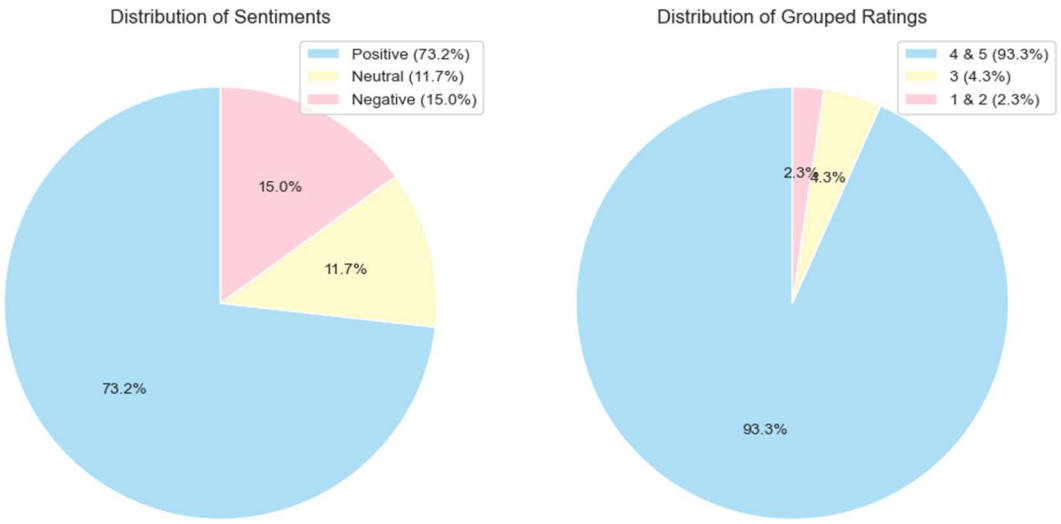

Figure 10 – NCR Distribution of Sentiments and Distribution of Grouped Ratings

By checking if the sentiment assessed by the model matched the rating associated with each review, we concluded that the model made 20,316 correct predictions which means it achieved an accuracy of 69.88%[11].

Table 4 shows that the NCR model is most competent in detecting positive sentiments, achieving a high precision of 93.89%, but its recall stands at 73.68%, indicating it missed a substantial portion of actual positive reviews. Its performance in recognizing negative and neutral sentiments is markedly poor. For negative sentiments, even though the recall

[11] It was not possible to calculate the coverage for the NCR lexicon model because we were not able to access the full lexicon directly.

is relatively high at 27.46%, the precision is a mere 4.29%. Similarly, the model's neutral sentiment detection has a precision of just 4.03% and a recall of 10.94%.

	precision	recall	f1-score	support
Negative	4.29%	27.46%	7.41%	2.34%
Neutral	4.03%	10.94%	5.89%	4.33%
Positive	93.89%	73.68%	82.56%	93.33%
macro avg	34.07%	37.36%	31.96%	100.0%
weighted avg	87.9%	69.88%	77.48%	100.0%

Table 4 – NCR Metrics Report

Overall, the NCR lexicon method demonstrates a precision of 87.09%, a recall of 69.88%, and an F1-score of 77.48%. Even though the NCRlexicon method performed well, it performed the worst out of all the models tested.

4.1.4 SentiWordNet

The SentiWordNet model first tokenizes the text, lemmatizes the tokens, and then matches them to their respective synsets. By accumulating the positive and negative scores from these synsets, a cumulative sentiment score is achieved. This score is then mapped to the categories 'Positive', 'Negative', or 'Neutral', depending on whether the score exceeds, falls short of, or equals zero, respectively.

Figure 11 shows that the 'Positive' sentiment dominates yet discernible discrepancies are evident. SentiWordNet classified 78.4% of the reviews as positive, in contrast to the actual 93.3%. Additionally, it miscategorized several reviews into 'Negative' and 'Neutral' sentiments.

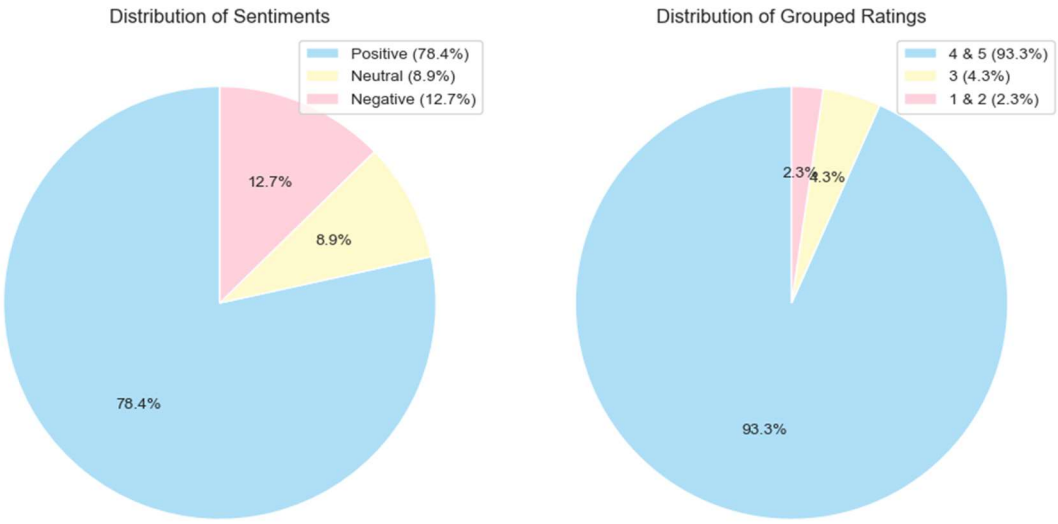

Figure 11 – SentiWordNet Distribution of Sentiments and Distribution of Grouped Ratings

By analysing whether the sentiment determined by the SentiWordNet model aligned with the given rating for each review, it was ascertained that the model's predictions were correct for a significant portion of the reviews - It attained an accuracy of 75.53%.

The coverage was calculated by determining the proportion of tokens in the dataset that are also present in the SentiWordNet lexicon. A coverage of 55.75% indicates that SentiWordNet contains slightly more than half of the unique words present in the dataset.

Table 5 shows that the model has a high precision of 94.69% for the 'Positive' category, but relatively low precision values for 'Negative' and 'Neutral' at 6.13% and 5.49%, respectively. The recall values suggest that the model correctly identifies 79.58% of the positive reviews but overestimates the 'Negative' and 'Neutral' sentiments with recall rates of 33.13% and 11.27% respectively. The F1 scores further validate these findings, with the model performing best for the 'Positive' sentiment at 86.48%.

	precision	recall	f1-score	support
Negative	6.13%	33.13%	10.34%	2.34%
Neutral	5.49%	11.27%	7.39%	4.33%
Positive	94.69%	79.58%	86.48%	93.33%
macro avg	35.44%	41.33%	34.74%	100.0%
weighted avg	88.75%	75.53%	81.27%	100.0%

Table 5 – SentiWordNet Metrics Report

The SentiWordNet model's overall performance boasts a precision of 88.75%, a recall of 75.53% and an F1 score of 81.27%. This model exhibits a commendable performance and exceeds our expectations, given that the studies we analysed got an accuracy of 68%.

4.1.5 Pattern

The Pattern sentiment analysis model offers a polarity score for a given text, ranging between -1 and 1. Based on this score, each review was classified as "Positive" if the polarity was greater than 0, "Negative" if it was less than 0, and "Neutral" if it equalled 0.

Figure 12 displays this distribution of the sentiments and the grouped ratings. It becomes evident that the sentiment distribution determined by the model closely mirrors the genuine sentiments expressed in the reviews.

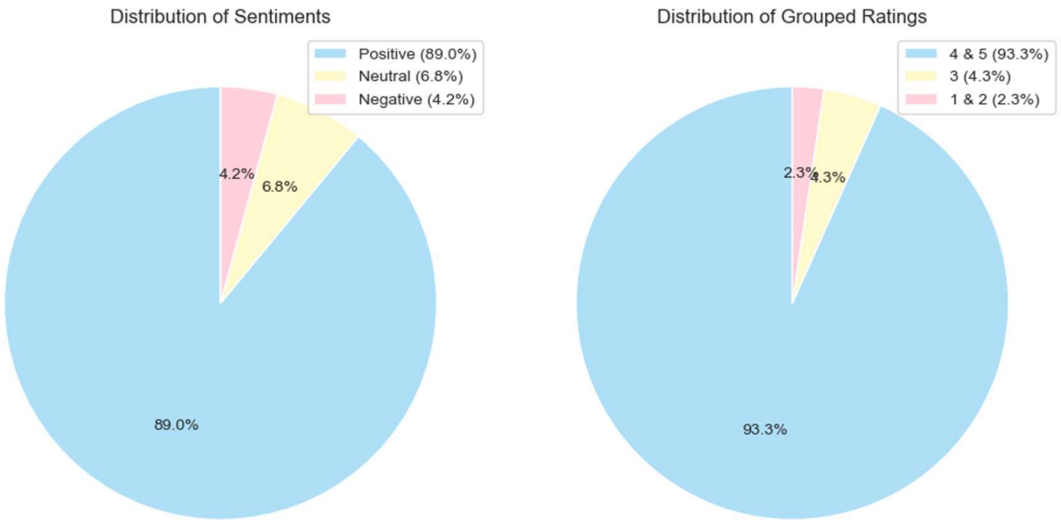

Figure 12 – Pattern Distribution of Sentiments and Distribution of Grouped Ratings

The model's accuracy was assessed by determining if the sentiment deduced by the model matches the associated review rating. From the total reviews, the model made 24,851 correct predictions, whereas 4,160 were incorrect which implies that the model's accuracy was 85.1%[12].

[12] Much like our experience with the TextBlob model, the Pattern's lexicon is not openly accessible and therefore it was not possible to calculate the model's coverage.

Regarding the model's ability to assign individual sentiments, it achieved high precision (94.51%) for positive reviews, but relatively lower values for negative (14.79%) and neutral (5.21%). The recall suggests the model captured 90.13% of actual positive sentiments, but only around 27% of the negative and 8% of the neutral ones. The F1-score, stands at 92.72% for positive sentiments, showcasing the model's strength in recognizing positive reviews. However, it's less proficient with neutral and negative reviews, as indicated by F1 scores of 6.35% and 19.04%, respectively (Table 6).

	precision	recall	f1-score	support
Negative	14.79%	26.72%	19.04%	2.34%
Neutral	5.21%	8.14%	6.35%	4.33%
Positive	94.51%	90.13%	92.27%	93.33%
macro avg	38.17%	41.66%	39.22%	100.0%
weighted avg	88.78%	85.1%	86.83%	100.0%

Table 6 – Pattern Metrics Report

Lastly, we calculated the weighted average of the performance metrics for the Pattern model's sentiment predictions achieving a precision of 88.78%, a recall of 85.01%, and an F1-score of 86.83%.

4.1.6 AFINN

The AFINN sentiment analysis model calculates a cumulative sentiment score based on the individual scores of words present in the review. This cumulative score is then used to categorize the sentiment of each review: scores above zero are classified as 'Positive', scores below zero are 'Negative', and a score of zero indicates a 'Neutral' sentiment.

The distribution of the sentiments assigned to each review is presented in Figure 13 When compared to the distribution of the grouped ratings, we can observe that both graphs have similar distributions, but the model categorized as 'neutral' and 'negative' has a higher number of reviews.

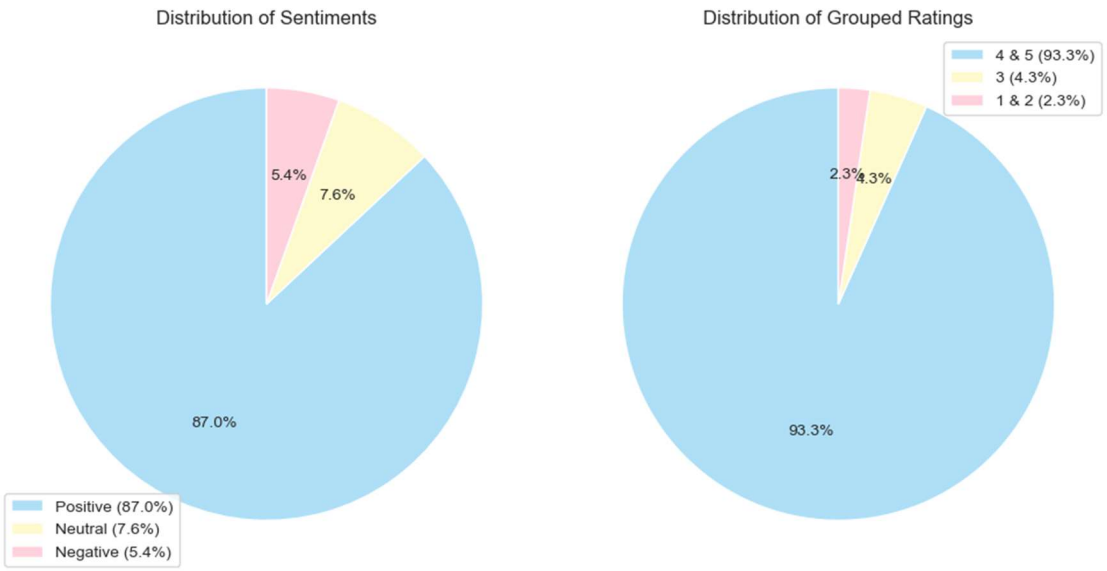

Figure 13 – AFINN Distribution of Sentiments and Distribution of Grouped Ratings

The sentiments predicted by the model were matched with the associated ratings to determine their correctness. Out of the total reviews, the model made 24,126 accurate predictions, leading to an overall accuracy of 84.11%.

Only 10.84% of the unique tokens from our dataset were recognized and scored by the AFINN model, suggesting that a significant portion of our dataset's vocabulary isn't directly accounted for in the AFINN sentiment lexicon[13].

Table 7 shows a strong performance predominantly for 'Positive' reviews, achieving a precision of 95.08% and recall of 88.6%. However, the model had weaker performance for 'Negative' and 'Neutral' reviews, with precisions of 13.06% and 9.4% respectively, though it exhibited a relatively higher recall for the 'Negative' category at 30.3%.

[13] Since the AFINN's coverage was calculated by determining which tokens from our data received a non-zero score from AFINN and therefore were present in AFINN's lexicon, the 10.84% is an approximation of the model's coverage because there might be words that the model assigned a zero score meaning the word had a neutral sentiment.

	precision	recall	f1-score	support
Negative	13.06%	30.3%	18.26%	2.34%
Neutral	9.4%	16.48%	11.97%	4.33%
Positive	95.08%	88.6%	91.73%	93.33%
macro avg	39.18%	45.13%	40.65%	100.0%
weighted avg	89.45%	84.11%	86.55%	100.0%

Table 7 – AFINN Metrics Report

The AFINN sentiment analysis model exhibited a weighted precision of 89.45%, a recall of 84.11%, and an F1-score of 86.55%. These metrics reflect a good performance.

4.1.7 Opinion Lexicon

Bing Liu's opinion lexicon tokenized each review and then cross-referenced with its lexicon, which consists of distinct sets of positive and negative words. For each review, a wordcount was performed for both the positive and negative word sets.

The sentiment of the review was determined based on the relative counts of positive and negative words. If a review had more (less) positive than negative words, it was categorized as "Positive" (“Negative”). In instances where the positive and negative word counts were equal, the review was labelled as "Neutral".

The model classified 30,031 reviews as 'Positive', which represents 86.7% of all reviews. This distribution is depicted in Figure 14 which reveals a striking similarity between the model's sentiment allocation and the authentic sentiments present in the reviews.

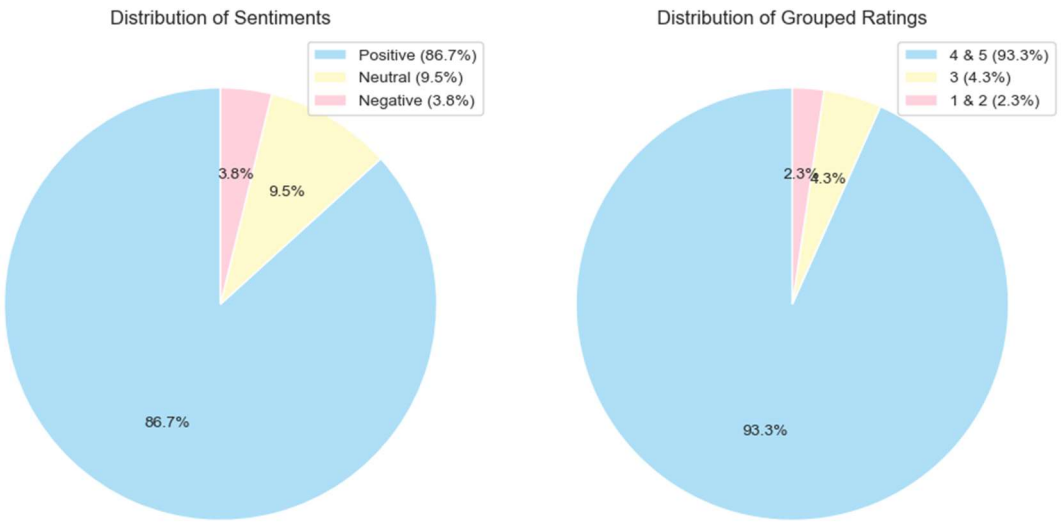

Figure 14 – Opinion Lexicon Distribution of Sentiments and Distribution of Grouped Ratings

By comparing the predictions against the actual ratings, it was determined how accurately the model's classifications matched the real sentiments. This led to an overall accuracy rate of 84.23%.

The unique words from the reviews were compared with Bing Liu's combined lexicon it was discovered that 15.98% of the unique words from the reviews were present in the lexicon. This indicates that most terms in the reviews are not covered by the lexicon.

The model showed a strong performance in identifying positive reviews with a precision of 95.36% and recall of 88.62%. However, it exhibited lower precision and recall for neutral and negative sentiments. For negative reviews, it achieved a precision of 16.38% and recall of 26.6%, while for neutral reviews, the precision was 9.52% and recall was 20.81%. This suggests that while the model is adept at pinpointing positive sentiments, it tends to misclassify a notable proportion of neutral and negative reviews (Table 8).

	precision	recall	f1-score	support
Negative	16.38%	26.6%	20.27%	2.34%
Neutral	9.52%	20.81%	13.07%	4.33%
Positive	95.36%	88.62%	91.87%	93.33%
macro avg	40.42%	45.34%	41.73%	100.0%
weighted avg	89.79%	84.23%	86.78%	100.0%

Table 8 – Opinion léxicon Metrics Report

In conclusion, the weighted performance measures were calculated - precision of 89.79%, the recall was 84.23% and the F1-score stood at 86.78%.

4.1.8 LabMT

The LabMT uses valence scores for words, which indicate the emotional sentiment of each word and used them to classify the reviews into corresponding sentiment labels. Reviews with scores exceeding 5.5 were labelled 'Positive', between 5.0 and 5.5 conveyed a neutral stance, earning the 'Neutral' tag and below 5.0 exuded negativity, leading to a 'Negative' classification.

The model identified 31,591 reviews as 'Positive', accounting for 91.2% of the total. Moreover, the model's classification showcased a higher count of 'Neutral' reviews than the actual tally. This comparison is captured in Figure 15.

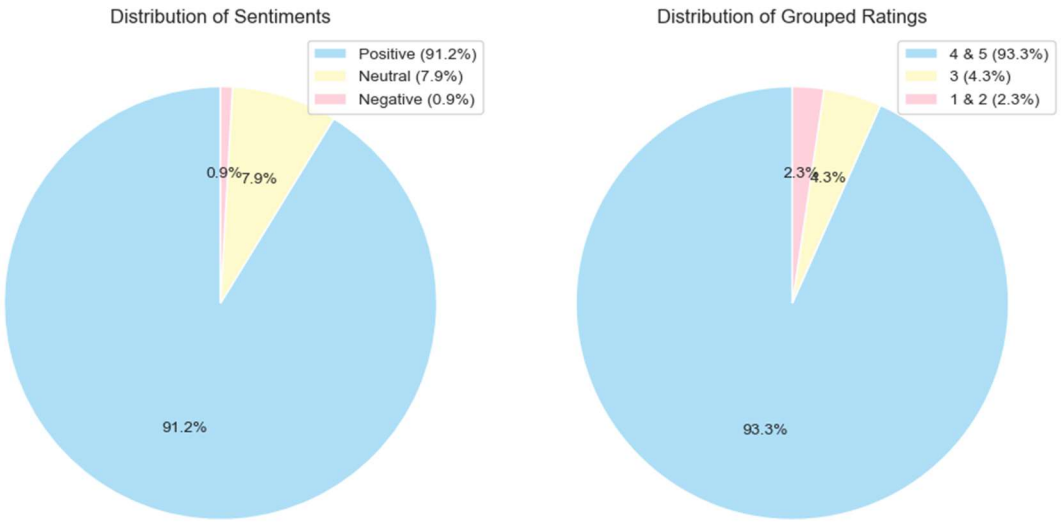

Figure 15 – LabMT Distribution of Sentiments and Distribution of Grouped Ratings

The LabMT model was applied to our dataset to predict the sentiment of each review. Out of the total reviews, 30,247 were accurately predicted which translates to an impressive accuracy rate of 87.35%.

Comparing the words in our dataset with the LabMT lexicon, it was discerned that approximately 49.26% of the unique words from the reviews were represented in the LabMT dictionary.

The model performs exceptionally well for 'Positive' reviews, achieving a precision of 94.72% and a recall of 92.59%. However, for the 'Negative' and 'Neutral' categories, the performance is not good with precisions of 14.29% and 10.3%, and recalls of 5.42% and 18.75%, respectively (Table 9).

	precision	recall	f1-score	support
Negative	14.29%	5.42%	7.86%	2.34%
Neutral	10.3%	18.75%	13.3%	4.33%
Positive	94.72%	92.59%	93.64%	93.33%
macro avg	39.77%	38.92%	38.26%	100.0%
weighted avg	89.18%	87.35%	88.15%	100.0%

Table 9 – LabMT Metrics Report

The sentiment classification attained a precision of 89.18%, coupled with a recall of 87.35% and the harmonized measure, the F1-score, stood at 88.15%.

4.1.9 ANEW

To implement the ANEW model, we computed the mean scores for valence, arousal, and dominance by breaking down each review into individual words. The mean of these scores is computed and then we calculate the average of the three mentioned scores and get the ‘mean_score’ variable which effectively amalgamates the sentiments of the reviews into a singular metric.

The reviews were then classified based on their 'mean score'. Reviews boasting a score exceeding 4 as 'Positive' l, between 3 and 4 as 'Neutral' and scoring less than 3 as 'Negative'.

Figure 16 – ANEW Distribution of Sentiments and Distribution of Grouped Ratings

From a glance at Figure 16, it is evident that the overwhelming majority of reviews are categorized as 'Positive'. A mere 134 reviews were labelled 'negative', constituting 0.4% of all reviews and only 11 reviews fell into the 'neutral' category, rounding off to 0%. This sentiment distribution stands in contrast to the patterns observed in the second graph, which represents the distribution of grouped ratings.

The sentiments predicted by the model were matched with the associated ratings to determine their correctness and the model achieves an impressive accuracy rate of 92.94%.

The unique tokens from the dataset were cross-referenced with the ANEW lexicon to identify which of them existed within the lexicon. The findings revealed that 43.01% of the unique tokens in the reviews were present in the ANEW lexicon.

Table 10 reveals that, even though the model performs well for the 'Positive' class, with a precision of 93.32% and a recall of 99.58%, it struggles to accurately classify the 'Negative' and 'Neutral' classes, with near-zero values for precision and recall.

	precision	recall	f1-score	support
Negative	2.99%	0.49%	0.85%	2.34%
Neutral	0.0%	0.0%	0.0%	4.33%
Positive	93.32%	99.58%	96.35%	93.33%
macro avg	32.1%	33.36%	32.4%	100.0%
weighted avg	87.17%	92.94%	89.94%	100.0%

Table 10 – ANEW Metrics Report

The model exhibited an admirable performance with an overall precision of 87.17%, a recall of 92.94% and the F1-score stands at 89.94%.

4.3 Comparison

Table 11 presents the metrics we calculated on the 9 various sentiment analysis models:

Model	Accuracy	Precision	Recall	F1-score	Coverage
Vader	86.55%	89.21%	86.55%	87.77%	13.28%
TextBlob	85.13%	88.76%	85.13%	86.84%	
NCRLex	69.88%	87.90%	69.88%	77.48%	
SentiWordNet	75.53%	88.75%	75.53%	81.27%	55.75%
Pattern	85.10%	88.78%	85.10%	86.83%	
AFINN	84.11%	89.45%	84.11%	86.55%	10.84%
Bing Liu's	84.23%	89.79%	84.23%	86.78%	15.98%
LabMT	87.35%	89.18%	87.35%	88.15%	49.26%
ANEW	92.94%	87.17%	92.94%	89.94%	43.01%

Table 11 – Model's Metrics

Starting with Accuracy, the ANEW model clearly stands out with the highest accuracy at 92.94%. This indicates that ANEW correctly classified the sentiments in the given dataset more often than any other model. Conversely, NCRLex seems to struggle the most in this aspect with an accuracy of just 69.88%.

Bing Liu's model excels with the highest precision of 89.79%, suggesting that when it classifies a review, it's most likely correct. The lowest precision is observed in ANEW, despite its high accuracy, indicating it has a higher rate of misclassifying reviews.

Regarding the recall metric, ANEW triumphs again with a recall of 92.94%, which means it is very efficient at capturing the reviews' class without mistakenly classifying them. NCRLex, with a recall of 69.88%, is at the other end of the spectrum, suggesting it might be missing a significant portion of reviews.

The ANEW model leads in the F1-score category as well, with an impressive score of 89.94%, reinforcing its overall robustness. NCRLex manages a relatively decent F1-score of 77.48% but is still the lowest when compared to the other models.

However, as mentioned previously, the precision, recall and F1-score assigned to each model were calculated using the weighted average of the precision and recall of the individual classes because the dataset is skewed towards the 'Positive' reviews. It's crucial to highlight that, while the ANEW model showcases the best results in terms of overall accuracy, recall, and F1 score, its performance in detecting and classifying 'Neutral' and 'Negative' reviews was subpar. It registered the poorest outcomes among all models in precision, recall, and F1-score for these two classes. Conversely, Bing Liu's lexicon demonstrated superior performance in classifying all three categories and, had the macro-average been employed instead of the weighted average, this model would've had the best results in overall precision, recall, and F1-score.

Regarding the last metric in table 11, even though we couldn't calculate the coverage for all models, we concluded that SentiWordNet and LabMT show substantial coverage, with 55.75% and 49.26% while models like AFINN have relatively lower coverage (10.84%).

We further evaluate the consistency among the models by calculating their agreement. Table 12 illustrates the percentage of reviews for which two distinct models reached the

	Vader	TexBlob	NCRLex	SentiWordNet	Pattern	AFINN	Bing Lius	LabMT	ANEW
Vader	100.0	87.86	72.65	77.78	87.93	92.42	87.67	87.55	89.92
TexBlob		100.0	71.01	76.39	99.85	86.65	86.09	85.42	88.97
NCRLex			100.0	66.34	70.97	72.46	69.84	71.35	73.15
SentiWordNet				100.0	76.31	76.39	75.35	75.81	78.27
Pattern					100.0	86.54	86.02	85.36	88.90
AFINN						100.0	87.46	86.17	86.87
Bing Lius							100.0	85.8	86.54
LabMT								100.0	91.04
ANEW									100.00

same sentiment conclusion, highlighting the degree of consensus or divergence in their classifications.

The agreement table reveals patterns of concurrence among the sentiment models. TextBlob and Pattern stand out with an almost identical classification pattern, displaying an impressive 99.85% agreement. This can be justified by the fact that both models derive from the pattern.en library and therefore used very similar techniques to classify the reviews. Other models exhibit a strong alignment such as Vader and AFINN with a 92.42% agreement, and LabMT and ANEW with a 91.04% agreement, even though they adopt very different techniques. In contrast, NCRLex frequently diverges from other models, even with LabMT which was also constructed with the MTurk.

In summary, while the ANEW model outperforms the others in most categories, it has its limitations, especially in precision and potentially detecting negative and neutral sentiments. Bing Liu's model achieved the highest precision and SentiWordNet and LabMT had the most extensive coverage.

Chapter 5: Conclusions and Limitations

5.1 Overall Conclusions

As we ventured into the intricate realms of sentiment analysis, we provided an extensive background on its history, evolution, and significance, particularly emphasizing the relevance of our chosen theme.

Drawing from the literature review on the subject, we proceed to the methodology and implemented various models on our chosen dataset. Their performances were evaluated, dissected, and contrasted.

In the subsequent sections of this chapter, we will condense our research's conclusions, elucidate the inherent limitations faced, and outline the potential research for future exploration. This project aims to enrich the academic discourse on sentiment analysis and make a lasting contribution to this dynamic field.

The findings of this project might have substantial implications for various stakeholders. For businesses, particularly Amazon, understanding customer sentiment at scale could empower data-driven decision-making, enabling the tailoring of marketing strategies, product improvements, and customer service interventions based on detailed sentiment insights.

For developers and data scientists, the results may provide valuable benchmarks and insights into the effectiveness of lexicon-based sentiment analysis approaches, inspiring further refinement of these tools and methods. The findings could potentially contribute to the design of more advanced, efficient, and accurate sentiment analysis methods, including methods combining Lexicon-based approaches with Machine Learning.

Lastly, the findings could lay a foundation for future research, inviting scholars to explore deeper into the intricacies of online review sentiment analysis and explore the comparative strengths and weaknesses of the different approaches used in this project.

Our primary objective is to contribute to the study of sentiment analysis. Our exploration showcased that sentiment analysis holds substantial promise in predicting Amazon ratings. The implemented models were able to discern underlying sentiments with reasonable accuracy.

Secondly, it was also important to evaluate and compare the performance of the various lexicon-based models in predicting sentiments of Amazon reviews. The implemented models varied in their performance, with the ANEW model showing superior metrics in accuracy, recall, and F1 score. However, its shortcomings in identifying neutral and negative reviews were evident.

This research journey into sentiment analysis has reached several pivotal conclusions. The ease of implementation of the chosen sentiment analysis models stands out prominently. Not only were these models easy to implement, but they also provided results rapidly. Additionally, they're available to all and do not have any associated costs.

In terms of performance, the models explored in this research managed to put forth admirable results. All models presented accuracies above 69% which emphasized their sentiment extraction capability.

Diving deeper into model-specific performances, the NCR model had certain limitations. Its foundation on an adjective dictionary might have caused it to falter with short-text reviews that lacked adjectives. Moreover, the model's intrinsic design to classify the text into nine categories, as opposed to the three used in this research, possibly compromised its adaptability and accuracy.

On the other hand, the ANEW model has been a star performer in this research, it achieved the best metrics across accuracy, recall, and F1 score, and its coverage is considerably high too. However, a critical observation is its relatively weaker performance in detecting and categorizing 'Neutral' and 'Negative' reviews. Given the dataset's positive skew, it remains an open question on how ANEW would perform on a more balanced dataset.

In conclusion, this project demonstrates the immense potential of lexicon-based models in sentiment analysis. While they might not attain perfection, their combination of speed, accuracy, and accessibility makes them invaluable tools in the realm of data analytics.

5.4 Limitations

One of the most significant limitations faced during this project was the skewed nature of our dataset. Most of the reviews in our dataset were rated '4' and '5' ('Positive). This skewness poses a challenge, as models could not be properly evaluated on their ability to

detect and categorize neutral and negative sentiments due to the sheer volume of positive reviews. Hence, the efficiency of a model might be interpreted based solely on its ability to recognize positive sentiments and therefore the calculated metrics might not reflect the model's true capabilities.

The amount of sentiment analysis tools is vast and varied. During the project's course, it was observed that certain models, which might have potentially outperformed the ones chosen for our analysis, were inaccessible due to licensing restrictions.

Another limitation was our inability to calculate the coverage for some models. Without this metric for certain models, we lose insight into a crucial aspect of their performance.

Addressing these limitations in subsequent research could lead to an even clearer understanding of sentiment analysis tools and their applicability in real-world scenarios.

5.5 Recommendations for Future Work

In the pursuit of improving sentiment analysis for predicting Amazon ratings, a future step could involve closely examining the incorrect predictions made by each model. By scrutinizing these misclassifications, gain a deeper understanding of the models' inherent weaknesses and the specific scenarios where they falter.

Furthermore, it would be highly beneficial to test their performance on a more balanced dataset. This will provide a clearer picture of their ability to accurately detect negative and neutral sentiments.

Another way that would complement this project would be to evaluate a wider spectrum of sentiment analysis models against the current dataset. By assessing a wider range of models, extract deeper insights into the nuances of sentiment analysis and its applicability to e-commerce review systems.

Beyond just expanding the lexicon-based, a truly promising area of exploration lies in the domain of machine learning, deploying sophisticated models, such as recurrent neural networks (RNNs) on the same dataset and comparing the results. One would assume that machine learning models would outperform their rule-based counterparts in predicting Amazon ratings. However, the question would be how much and whether would it be worth the extra time and computational needs.

References

Abiola, O., Abayomi-Alli, A., Tale, O. A., Misra, S., & Abayomi-Alli, O. (2023). Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser. *Journal of Electrical Systems and Information Technology*, *10*(1). https://doi.org/10.1186/s43067-023-00070-9

Akkaya, C., Conrad, A., Wiebe, J., & Mihalcea, R. (2010). *Amazon Mechanical Turk for Subjectivity Word Sense Disambiguation*. http://mturk.amazon.com

Aljuhani, S. A., & Alghamdi, N. S. (2019). A comparison of sentiment analysis methods on Amazon reviews of Mobile Phones. *International Journal of Advanced Computer Science and Applications*, *10*(6), 608–617. https://doi.org/10.14569/ijacsa.2019.0100678

Andrea, A. D. ', Ferri, F., & Grifoni, P. (2015). Approaches, Tools and Applications for Sentiment Analysis Implementation. In *International Journal of Computer Applications* (Vol. 125, Issue 3). http://messenger.yahoo.com/features/emoticons

Ashok Kumar, J., & Abirami, S. (2015). An Experimental Study of Feature Extraction Techniques in Opinion Mining. *International Journal on Soft Computing, Artificial Intelligence and Applications*, *4*(1), 15–21. https://doi.org/10.5121/ijscai.2015.4102

Baid, P., Gupta, A., & Chaplot, N. (2017). Sentiment Analysis of Movie Reviews using Machine Learning Techniques. In *International Journal of Computer Applications* (Vol. 179, Issue 7). http://reviews.imdb.com/Reviews

Bradley, M. M., & Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*.

Datafiniti. (2019). *Consumer Reviews of Amazon Products*. Https://Www.Kaggle.Com/Datasets/Datafiniti/Consumer-Reviews-of-Amazon-Products.

De Smedt, T., & Daelemans, W. (2012). Pattern for Python. In *Journal of Machine Learning Research* (Vol. 13). http://www.clips.ua.ac.be/pages/pattern.

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing*, *34*(6), 480–488. https://doi.org/10.1108/JCM-03-2017-2141

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, *6*(12). https://doi.org/10.1371/journal.pone.0026752

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, *2*(1). https://doi.org/10.1186/s40537-015-0015-2

Filho, P. P. B., & Pardo, T. A. S. (2013). *NILC USP: A Hybrid System for Sentiment Analysis in Twitter Messages* (Vol. 2). http://github.com/pedrobalage/SemevalTwitterHybridClassifier

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. *COSN 2013 - Proceedings of the 2013 Conference on Online Social Networks*, 27–37. https://doi.org/10.1145/2512938.2512951

Hamouda, A., & Rohaim, M. (2011). Reviews Classification Using SentiWordNet Lexicon. In *The Online Journal on Computer Science and Information Technology (OJCSIT)* (Issue 2). https://www.researchgate.net/publication/267249616

Hardeniya, T., & Borikar, D. A. (2016). Dictionary Based Approach to Sentiment Analysis-A Review. *International Journal of Advanced Engineering, Management and Science (IJAEMS)*, *2*(5). www.ijaems.com

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing*, *40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Hirschberg, J., & Manning, C. D. (2015). *Advances in natural language processing*. www.sciencemag.org

Hu, M., & Liu, B. (2004). *Mining and Summarizing Customer Reviews*.

Hutto, C. J., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. http://sentic.net/

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment Analysis of Short Informal Texts. In *Journal of Artificial Intelligence Research* (Vol. 50). http://www.cs.york.ac.uk/semeval-2013/task2

Koto, F., & Adriani, M. (2015). A Comparative Study on Twitter Sentiment Analysis: Which Features are Good? In *Natural Language Processing and Information Systems* (pp. 453–457). http://www.springer.com/series/7409

Liu, B. (2010). *Sentiment analysis and subjectivity*. https://www.researchgate.net/publication/228667268

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018a). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. In *Computer Science Review* (Vol. 27, pp. 16–32). Elsevier Ireland Ltd. https://doi.org/10.1016/j.cosrev.2017.10.002

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018b). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. In *Computer Science Review* (Vol. 27, pp. 16–32). Elsevier Ireland Ltd. https://doi.org/10.1016/j.cosrev.2017.10.002

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The Geography of Happiness: Connecting Twitter Sentiment and Expression,

Demographics, and Objective Characteristics of Place. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0064417

Mohammad, S. M., & Turney, P. D. (2010). *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon*. http://www.wjh.harvard.edu/

Nielsen, F. Å. (2011). *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs*. http://arxiv.org/abs/1103.2903

Ohana, B., Tierney, B., & Brendan, T. (2009). Sentiment Classification of Reviews Using SentiWordNet. In *Computer Sciences*. https://arrow.tudublin.ie/scschcomconhttps://arrow.tudublin.ie/scschcomcon/293

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment Classification using Machine Learning Techniques*. EMNLP. http://www.cs.cornell.edu/people/pabo/movie-review-data/.

Pano, T., & Kashef, R. (2020). A complete vader-based sentiment analysis of bitcoin (BTC) tweets during the ERA of COVID-19. *Big Data and Cognitive Computing*, *4*(4), 1–17. https://doi.org/10.3390/bdcc4040033

Parashar, M., Jaypee Institute of Information Technology University, University of Florida. College of Engineering, Institute of Electrical and Electronics Engineers. Delhi Section, & Institute of Electrical and Electronics Engineers. (2013). *Big Data: Issues, Challenges, Tools and Good Practices*.

Patel, D. (2015). *Approaches for Sentiment Analysis on Twitter: A State-of-Art study*.

Pedrycz, W., Chen, S.-M., & Environment, A. (2016). *Studies in Computational Intelligence 639 Sentiment Analysis and Ontology Engineering*. http://www.springer.com/series/7092

Prager, J. (2006). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, *1*(2), 91–233. https://doi.org/10.1561/1500000001

Praveen Gujjar, A. P., & Prasanna Kumar, H. R. (2021). Sentiment Analysis:Textblob For Decision Making. *International Journal of Scientific Research & Engineering Trends*, *7*(2), 2395–2566. https://doi.org/10.1109/MDM.2013

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(1). https://doi.org/10.1140/epjds/s13688-016-0085-1

Saif M, M. (2017). *A Practical Guide to Sentiment Analysis*. http://www.springer.com/series/13199

Sebastiani, F., & Esuli, A. (2016). *SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining*. https://www.researchgate.net/publication/200044289

Shrestha, N., & Nasoz, F. (2019). Deep Learning Sentiment Analysis of Amazon.Com Reviews and Ratings. *International Journal on Soft Computing, Artificial*

Intelligence and Applications, *8*(1), 01–15. https://doi.org/10.5121/ijscai.2019.8101

Sohangir, S., Petty, N., & Wang, Di. (2018). Financial Sentiment Lexicon Analysis. *Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018*, *2018-January*, 286–289. https://doi.org/10.1109/ICSC.2018.00052

Sri, R., Niharika, Ch., Maneesh, K., & Ismail, Dr. M. (2019). Sentiment Analysis of Patients' Opinions in Healthcare using Lexicon-based Method. *International Journal of Engineering and Advanced Technology*, *9*(1), 6977–6981. https://doi.org/10.35940/ijeat.A2141.109119

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, *36*, 10–25. https://doi.org/10.1016/j.inffus.2016.10.004

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). *Lexicon-Based Methods for Sentiment Analysis*. http://direct.mit.edu/coli/article-pdf/37/2/267/1798865/coli_a_00049.pdf

Tanjim Ul Haque, Nudrat Nawal Saber, & Faisal Muhammad Shah. (2018). *Sentiment Analysis on Large Scale Amazon Product Reviews*.

Zad, S., Jimenez, J., & Finlayson, M. A. (2021). *Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon*. 102–113. https://doi.org/10.34703/gzx1-9v95/PO3YGX