SENTIMENT ANALYSIS FOR PRODUCT REVIEWS

NATURAL LANGUAGE PROCESSING PROJECT REPORT

Submitted by

ANANTH SHYAM S

22011101012

AVINASH M

22011101019

SEMESTER VI

BACHELOR OF TECHNOLOGY IN ARTIFICIAL INTELLIGENCE & DATA SCIENCE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SHIV NADAR UNIVERSITY CHENNAI

APRIL 2025

SHIV NADAR
UNIVERSITY
CHENNAI

TABLE OF CONTENTS

ABSTRACT

This project explores the domain of Sentiment Analysis for Product Reviews, a crucial task in Natural Language Processing (NLP) that involves determining the sentiment —positive or negative—expressed in textual customer feedback. Product reviews, being highly diverse, unstructured, and often context-dependent, present unique challenges such as sarcasm, mixed sentiment, and syntactic ambiguity. The purpose of this study is to systematically analyze and implement a range of sentiment analysis techniques, starting from foundational text processing methods and advancing toward state-of-the-art transformer-based language models.

The workflow begins with classical approaches such as Bag of Words (BoW) and TF-IDF, which rely on frequency-based representations of text. These are followed by embedding-based techniques like Word2Vec, which incorporate contextual similarity. As limitations in capturing semantic depth and syntactic structure became evident, the project transitions into using advanced deep learning models including BERT, Sentence-BERT, and a fine-tuned BERT classifier. At each stage, the study carefully evaluates performance, identifies the types of errors each model makes, and investigates how those limitations can be addressed—either by transitioning to a more capable model or by integrating hybrid techniques like combining lexical rules with embeddings or applying syntactic parse trees to handle conjunctions and compound structures.

The overarching goal of this study is not only to improve sentiment prediction accuracy but also to understand the working mechanisms of various models, why certain models fail under specific linguistic conditions, and how to remedy these gaps using either linguistic insights or deeper contextual models. The dataset consists of real-world product reviews collected from e-commerce platforms, which offer a rich source of diverse expressions and sentiments. The final outcome includes a detailed comparative analysis, a robust understanding of sentiment modeling progression, and insights into practical applications for building more intelligent sentiment-aware systems.

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

Sentiment analysis is a subfield of Natural Language Processing (NLP) that focuses on identifying and categorizing opinions expressed in a piece of text—particularly to determine the writer's attitude as positive, negative, or neutral. It plays a vital role in understanding public opinion, customer satisfaction, and brand perception. With the explosive growth of user-generated content in the form of reviews, ratings, and feedback, sentiment analysis has become an essential tool for both businesses and researchers aiming to extract actionable insights from text data.

This project focuses specifically on sentiment analysis of product reviews, where customers express their satisfaction or dissatisfaction regarding a product. These reviews, being informal and user-generated, often contain noise, slang, sarcasm, and complex sentence structures, making it a challenging NLP task.

1.2 PROBLEM STATEMENT

Traditional sentiment analysis models often struggle with ambiguity, syntactic complexity, and the nuances of human expression. Frequency-based models like Bag of Words and TF-IDF lack semantic understanding, while rule-based models may fail to generalize across varied writing styles. Although deep learning models like Word2Vec improve upon these limitations by embedding semantic context, they still fall short in handling complex linguistic structures such as negation, conjunctions, and sarcasm.

Thus, this study aims to investigate:

- How each NLP technique performs on product review data.
- What types of errors each method is prone to.
- How these issues can be addressed through more advanced or hybrid techniques, including the use of syntactic parsing and transformer-based models like BERT.

1.3 OBJECTIVES

The main objective of this project is to implement and evaluate a wide range of sentiment analysis techniques, starting from basic NLP models and progressing toward advanced deep learning approaches. This includes understanding the strengths and weaknesses of each method, and closely examining the kinds of errors they tend to produce—whether due to context ambiguity, negation, or syntactic complexity. Another key goal is to explore and develop hybrid strategies to overcome these limitations, such as combining lexical resources like WordNet with rule-based models like Vader or applying syntactic parsing to enhance BERT's handling of conjunctions. Finally, the project aims to fine-tune a transformer-based model specifically for product review sentiment classification, ensuring it adapts well to the nuances and domain-specific language found in real-world customer feedback.

1.4 MOTIVATION

Product reviews are a rich source of consumer insight, and accurate sentiment analysis can provide companies with crucial feedback for product improvement, customer engagement, and marketing strategies. However, most real-world reviews contain linguistic challenges that require more than just keyword spotting or statistical correlations. The motivation behind this study is to not only build performant models but also to understand the *why* behind their success or failure—bridging the gap between classical NLP and modern transformer-based methods. Through this step-by-step comparative approach, the project aims to contribute a comprehensive perspective on sentiment modeling across multiple NLP paradigms.

CHAPTER 2
BACKGROUND

2.1 SENTIMENT ANALYSIS TECHNIQUES

Sentiment analysis, or opinion mining, is a key area in NLP focused on determining whether a piece of text expresses a positive, negative, or neutral sentiment. Early methods used rule-based systems and lexicons like Vader, which rely on predefined word lists and heuristics for handling negations and intensifiers. While these are simple and interpretable, they often fail to capture deeper context or subtle language cues.

To improve accuracy, machine learning models like Naive Bayes and Logistic Regression were applied using frequency-based techniques such as Bag of Words (BoW) and TF-IDF. These were followed by word embeddings like Word2Vec and GloVe, which added semantic understanding but still lacked contextual sensitivity. The real leap came with deep learning and transformer-based models like BERT and Sentence-BERT, which understand word meaning in context and handle complex linguistic patterns, making them state-of-the-art for sentiment classification.

2.2 OVERVIEW OF EXISTING TECHNIQUES

Sentiment analysis has seen the development of a wide range of techniques, from simple rule-based systems to advanced deep learning models. Each technique brings its own strengths and limitations depending on the nature of the text and the domain of application. In this section, we outline several major approaches used in sentiment analysis, highlighting their functionality and relevant research where applicable.

2.2.1 VADER (Valence Aware Dictionary and sEntiment Reasoner)

VADER is a lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media. It combines a sentiment lexicon with heuristics to account for punctuation, capitalization, degree modifiers, and negations. While it is highly interpretable and lightweight, it often falls short in domain-specific tasks or when dealing with complex syntactic structures. [Hutto & Gilbert, 2014]

2.2.2 Bag of Words (BoW)

BoW is one of the most basic representations in NLP, converting text into a multiset of words, disregarding grammar and word order. Despite its simplicity, it remains a strong baseline and is often used in conjunction with classifiers like Naive Bayes or Logistic Regression. However, BoW fails to capture semantic meaning and syntactic structure.

2.2.3 TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF builds on BoW by down-weighting common words and up-weighting rare but potentially informative words. This makes it more effective in distinguishing documents based on keyword relevance. However, it still treats words independently and does not capture context. It remains popular in resource-constrained settings due to its efficiency.

2.2.4 Word2Vec

Introduced by Mikolov et al. (2013), Word2Vec learns word embeddings that capture semantic relationships by training on word co-occurrence. It allows for meaningful vector operations and captures similarity well. However, it produces static embeddings, meaning that word meaning does not change based on context.

2.2.5 LSTM and CNN-based Models

Recurrent neural networks like LSTMs (Long Short-Term Memory) and convolutional neural networks have been applied successfully to sentiment analysis, especially for modeling sequences and local features. They provide better

context handling than traditional models but require large datasets and substantial training time. [Zhang et al., 2015]

2.2.6 BERT (Bidirectional Encoder Representations from Transformers)

BERT, developed by Devlin et al. (2018), revolutionized NLP by enabling bidirectional context-aware word representations. Fine-tuning BERT on sentiment datasets leads to state-of-the-art results, particularly for nuanced and longer texts. It understands syntactic dependencies and subtle sentiment cues well.

2.2.7 RoBERTa and DistilBERT

RoBERTa, a robustly optimized variant of BERT, and DistilBERT, a lighter and faster version, offer performance and efficiency trade-offs. These models maintain BERT's strengths while catering to different deployment needs. RoBERTa achieves slightly higher accuracy while DistilBERT allows for faster inference.

2.3 FUNDAMENTALS OF SENTIMENT ANALYSIS

Sentiment analysis plays a crucial role in understanding customer perceptions and experiences, especially in the context of product reviews. These reviews are an invaluable source of consumer feedback and often influence purchasing decisions, brand reputation, and marketing strategies. By automatically determining whether a review expresses a positive or negative opinion, sentiment analysis allows businesses to quickly summarize large volumes of customer feedback and identify key areas of improvement. Beyond commercial utility, this analysis also provides insights into customer behavior, emotional trends, and expectations, making it a central component in modern consumer analytics.

2.3.1 Role of Sentiment in Customer Feedback Interpretation

Product reviews are typically unstructured, diverse in language, and rich with user opinion. Analyzing their sentiment helps in interpreting overall product reception and can uncover hidden dissatisfaction or appreciation. For example, a product with high ratings might still include negative sentiment within the review text, which would be missed by a purely numerical analysis. Sentiment detection can also highlight specific product features that customers love or dislike, allowing for more targeted improvements and marketing communication.

2.3.2 Sentiment Analysis Models in Product Review Context

In recent years, specialized language models have significantly advanced the field of sentiment analysis, particularly in the domain of product reviews. BERT (Bidirectional Encoder Representations from Transformers), a transformer-based model developed by Google, has been fine-tuned on various sentiment-labeled review datasets such as Amazon, Yelp, and IMDB, resulting in substantial performance improvements over traditional methods. These fine-tuned versions, including the standard bert-base-uncased and domain-adapted variants like those trained on the amazon_reviews_multi dataset, have demonstrated remarkable capability in capturing both contextual meaning and subtle syntactic structures.

Unlike frequency-based or lexicon-driven models, BERT can understand word meanings in context and adapt to the informal, diverse, and often nuanced language found in product reviews. When fine-tuned on clean, balanced datasets, BERT-based models consistently achieve sentiment classification accuracies in the range of 92–95%. This makes them especially suitable for applications requiring high precision and reliability, such as customer feedback systems and review-based recommender engines.

CHAPTER 3
METHODOLOGY

This chapter presents a comprehensive overview of the methodologies used to design, implement, and evaluate models for sentiment analysis. The process is illustrated in Fig 3.1, which provides a structured approach from data collection to model training and performance evaluation. Each methodological step is outlined in detail below, offering a clear understanding of the techniques and processes that enable effective sentiment analysis.

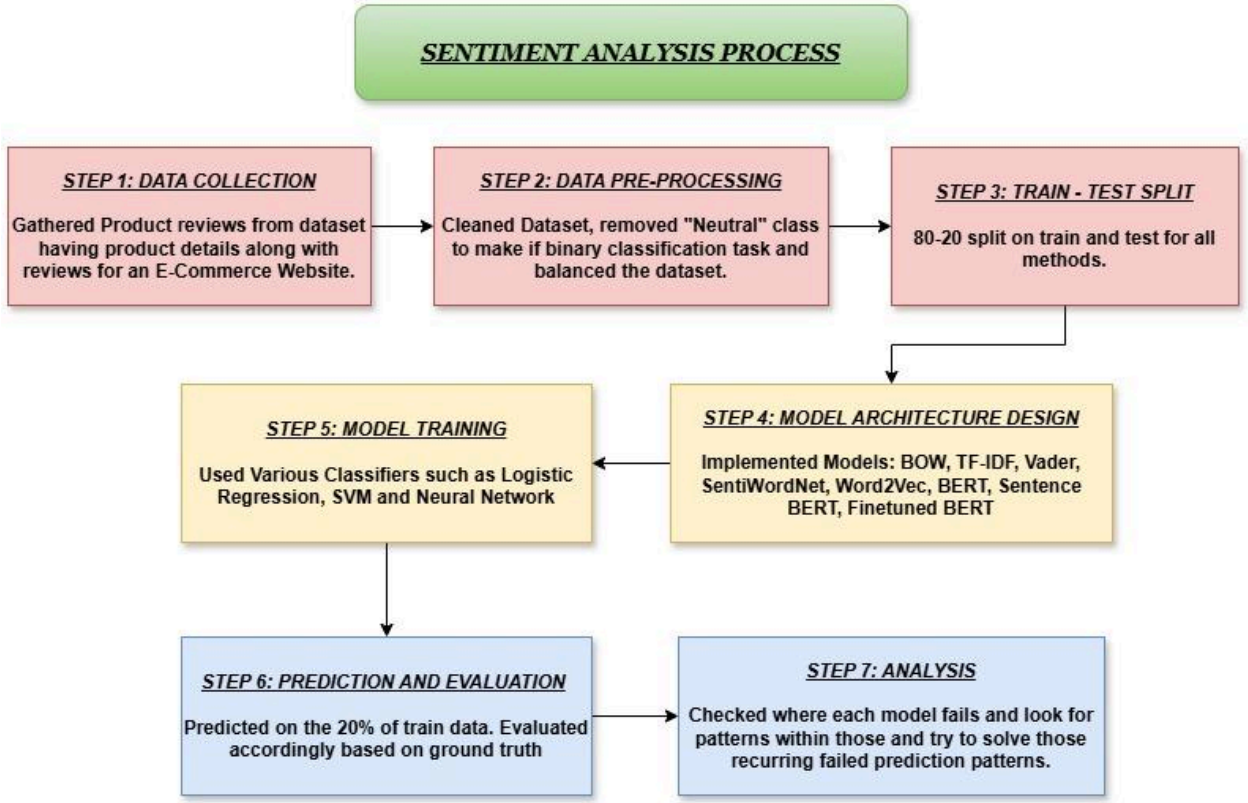

Figure 3.1

3.1 DATA COLLECTION

The dataset used in this project was sourced from Kaggle and contains customer reviews extracted from an e-commerce platform. Each entry in the dataset includes product details such as product name, category, rating, and a corresponding textual review written by a user. Additionally, each review is labeled with a sentiment tag—positive, negative, or neutral—based on the rating or predefined annotation. This labeled data serves as the ground truth for training and evaluating the sentiment analysis models. The dataset offers a rich variety of sentence structures, informal expressions, and product-related contexts, making it suitable for testing a wide spectrum of sentiment analysis techniques.

3.2 DATA PREPROCESSING

To prepare the dataset for effective sentiment classification, several preprocessing steps were undertaken. Initially, the dataset was cleaned by removing irrelevant columns, handling missing values, and standardizing text formats (e.g., lowercasing, removing special characters and HTML tags). The sentiment labels were simplified by removing the "Neutral" class, thereby converting the task into a binary classification problem with only "Positive" and "Negative" classes.

Since the dataset was imbalanced—having significantly more positive reviews than negative—sampling techniques were applied to balance the two classes. This was done by under-sampling the majority class to ensure an equal number of positive and negative reviews. This balance is crucial for training unbiased models and improving overall classification performance.

3.3 MODEL ARCHITECTURE AND DESIGN

This section elaborates on the various sentiment analysis model architectures implemented throughout the project. These models were selected for their ability to interpret textual data at different levels of complexity—from simple frequency-based representations to advanced contextual embeddings. To evaluate

effectiveness across a wide range of scenarios, we employed multiple classification algorithms, including Naive Bayes, Logistic Regression, Support Vector Machines (SVM), Random Forest, and Neural Networks.

Each model was assessed on its performance in classifying product review sentiments, particularly focusing on how well it handled challenges like negation, mixed sentiments, and syntactic nuances. This comparative study enabled us to better understand the strengths and limitations of each model and how they contribute to sentiment prediction accuracy.

3.3.1 Bag Of Words [BoW]

The Bag of Words model is one of the simplest and most widely used techniques in text classification tasks like sentiment analysis. It represents text data by converting each document into a fixed-length vector based on the frequency of words occurring in the document, disregarding grammar and word order. The matrix is generally very very sparse and may lead to overfitting. While BoW is computationally efficient and easy to implement, it fails to capture the contextual or semantic meaning of words, often leading to poor performance on complex language structures.

3.3.2 Term Frequency–Inverse Document Frequency (TF-IDF)

TF-IDF improves upon the Bag of Words model by not only considering how frequently a word appears in a document (Term Frequency), but also how rare or common it is across the entire corpus (Inverse Document Frequency). This helps in reducing the weight of common words that do not carry much semantic value and boosts the importance of distinctive terms. It is less sparse when compared to BoW but still generates a very sparse matrix. Though TF-IDF still lacks deep contextual understanding, it generally performs better than BoW in emphasizing meaningful words for sentiment classification.

3.3.3 VADER (Valence Aware Dictionary and sEntiment Reasoner)

VADER is a rule-based sentiment analysis tool specifically designed to analyze sentiments expressed in social media and short texts. It uses a sentiment lexicon where words are mapped to sentiment intensities and applies a set of grammatical and syntactic rules to account for factors like negations, punctuation, capitalization, and degree modifiers. VADER is particularly effective for quick analysis and interpretable results but tends to struggle with domain-specific language, sarcasm, or complex syntactic structures often found in detailed product reviews.

3.3.4 VADER + WordNet

To enhance VADER's capabilities in capturing deeper semantic meaning, we integrated it with WordNet, a lexical database of English. This hybrid approach expands the sentiment lexicon by including synonyms and related terms derived from WordNet, helping VADER handle a broader vocabulary, especially domain-specific words not originally present in its dictionary. Additionally, WordNet enables rudimentary sense disambiguation, allowing the model to better understand word usage in context. This combination improves VADER's coverage and effectiveness, especially in cases where reviews contain uncommon or descriptive language that standard VADER alone might misclassify.

3.3.5 Word2Vec

Word2Vec is an embedding-based technique that represents words as continuous vector values based on their surrounding context in a large corpus. Unlike frequency-based methods like BoW or TF-IDF, Word2Vec captures semantic relationships between words—for example, understanding that "great" and "excellent" have similar meanings. In this project, Word2Vec was used to convert reviews into dense vector representations, which were then fed into classifiers such as Logistic Regression or Neural Networks. This approach allows the model to generalize better across similar expressions and improves its ability to understand the overall sentiment of a review beyond exact word matches.

3.3.6 Word2Vec for Conjunctions (with Parse Tree Integration)

While Word2Vec captures semantic similarities between words, it often misses out on the syntactic structure of sentences—especially in cases with mixed or compound sentiments joined by conjunctions like "but", "although", or "however". To address this, we incorporated syntactic parsing using dependency parse trees to identify and isolate key sentiment phrases connected by such conjunctions.

For instance, in a review like *"The camera is great, but the battery life is disappointing"*, the model would treat the entire sentence uniformly without parsing. However, using dependency parsing, we can split the sentence into two clauses and separately analyze the sentiment for each clause. Word2Vec embeddings were then applied to each part, and rule-based logic (e.g., prioritizing the clause after "but") was used to refine the final sentiment classification. This hybrid approach improves accuracy in cases where sentiment is mixed or shifted mid-sentence, providing more context-sensitive predictions.

3.3.7. BERT

BERT (Bidirectional Encoder Representations from Transformers) represents a significant leap in Natural Language Processing by understanding the context of a word based on both its left and right surroundings. Unlike earlier models that process text in a unidirectional manner, BERT reads entire sequences at once, enabling it to capture complex linguistic patterns and contextual relationships.

For this project, we utilized the pre-trained “bert-base-uncased” model and fine-tuned it on the product review dataset. BERT’s contextual embeddings significantly outperformed traditional models, especially in handling nuanced expressions, sarcasm, and longer sentences. It also reduced the reliance on manual feature engineering, allowing the model to learn directly from raw text. As a result, BERT provided strong baseline performance for sentiment classification, particularly in ambiguous or context-heavy reviews.

3.3.8 BERT for Conjunctions (with Parse Tree Integration)

BERT, while powerful, can misinterpret sentences with mixed sentiments joined by conjunctions like *but* or *however*. To improve this, we used syntactic parsing to detect such conjunctions and split sentences into sentiment-bearing clauses. Each clause was analyzed individually using BERT, helping the model better understand contrasting opinions within the same review. This approach enhanced accuracy, especially in complex, compound sentences, and made predictions more aligned with human judgment.

3.3.9 Sentence BERT

Sentence-BERT (SBERT) is a modification of BERT that uses Siamese and triplet network structures to generate semantically meaningful sentence embeddings. Unlike traditional BERT, which is optimized for token-level tasks, SBERT is designed to compare entire sentences efficiently. In our project, SBERT was used to encode full review sentences into fixed-length vectors, enabling more accurate sentiment classification using classifiers like Logistic Regression or SVM. Its ability to capture sentence-level semantics helped improve performance on reviews with nuanced expressions.

3.3.10. Fine-Tuned BERT

Fine-tuned BERT involves taking a pre-trained BERT model—in our case, bert-base-uncased —and training it further on our domain-specific dataset of product reviews. This process helps the model better understand the contextual patterns, language nuances, and sentiment expressions specific to user-generated content in e-commerce. We added a classification head on top of BERT and fine-tuned the entire architecture using labeled data for binary sentiment classification. As a result, the model adapted well to real-world review data, capturing subtleties like sarcasm, negations, and comparative language. This approach consistently outperformed traditional and embedding-based models, demonstrating BERT's superior capability in handling both syntactic and semantic aspects of sentiment.

CHAPTER 4
RESULTS AND ANALYSIS

4.1 MODEL PERFORMANCE COMPARISONS

To evaluate the effectiveness of each sentiment analysis technique, we conducted a comprehensive comparison across all ten models implemented in this study. These ranged from basic frequency-based models like Bag of Words and TF-IDF to more advanced embedding and transformer-based approaches such as Word2Vec, Sentence-BERT, and Fine-Tuned BERT. Each model was assessed on a balanced binary sentiment dataset using accuracy, precision, recall, and F1-score. The analysis highlighted how performance steadily improved as we moved from simple lexical techniques to deep learning and contextual models, with the fine-tuned BERT model delivering the highest accuracy. This progression underscores the importance of contextual understanding and domain-specific fine-tuning in capturing nuanced sentiments present in real-world product reviews.

4.1.1 Bag Of Words [BoW]

```
Accuracy: 78.46%

Classification Report:
```

	precision	recall	f1-score	support
negative	0.99	0.57	0.72	26481
none	0.00	0.00	0.00	0
positive	0.83	0.98	0.90	29599
accuracy			0.78	56080
macro avg	0.61	0.52	0.54	56080
weighted avg	0.91	0.78	0.82	56080

The Bag of Words (BoW) model achieved an accuracy of 78.46%, but its limitations are evident in handling negations and contextual meaning. Although the precision for the negative class is high (0.99), the recall is low (0.57), indicating many negative reviews were misclassified. This is due to its frequency-based nature, where decisions are made solely on word counts, often misjudging phrases like "not good" as positive. The overall performance highlights BoW’s lack of depth in capturing nuanced language.

4.1.2 TF-IDF

```
Accuracy: 79.76%

Classification Report:
```

	precision	recall	f1-score	support
negative	0.99	0.61	0.76	26481
none	0.00	0.00	0.00	0
positive	0.89	0.96	0.92	29599
accuracy			0.80	56080
macro avg	0.62	0.53	0.56	56080
weighted avg	0.93	0.80	0.84	56080

The TF-IDF model achieved a slightly improved accuracy of 79.76%, indicating better handling of informative words compared to BoW. The model shows improved recall for the negative class (0.61) and strong performance for positive reviews (f1-score of 0.92), but still suffers from limitations in understanding negations and sentence structure. Since TF-IDF also treats words independently and lacks semantic understanding, it often misinterprets context, leading to incorrect classifications in nuanced or sarcastic reviews.

4.1.3 VADER

```
Accuracy: 88.51%

Classification Report:
```

	precision	recall	f1-score	support
negative	0.96	0.79	0.87	26481
positive	0.84	0.97	0.90	29599
accuracy			0.89	56080
macro avg	0.90	0.88	0.88	56080
weighted avg	0.89	0.89	0.88	56080

The Vader-based model shows a notable improvement with an accuracy of 88.51%, highlighting its strength in handling polarity using a sentiment lexicon and rule-based scoring. It performs well on both classes, especially positive reviews (recall of 0.97), due to its sensitivity to intensity modifiers and punctuation. However, it struggles with spelling variations, informal language, and mixed sentiments, often assigning a compound score of 0 when words are not found in its dictionary. Additionally, the absence of a fixed threshold in Vader makes classification inconsistent in borderline or compound statements.

4.1.4 VADER + WORDNET

```
Accuracy: 64.41%

Classification Report:

              precision    recall  f1-score   support

    negative       0.86      0.67      0.75      1751
    positive       0.27      0.52      0.35       401

    accuracy                           0.64      2152
   macro avg       0.56      0.59      0.55      2152
weighted avg       0.75      0.64      0.68      2152
```

This approach specifically targets the subset of reviews where Vader fails by assigning a compound score of 0, usually due to out-of-vocabulary words or spelling issues. With an accuracy of 64.41%, it demonstrates a notable improvement in handling these edge cases compared to standard Vader. The model recovers better precision for negative sentiment (0.86) and achieves a more balanced recall, addressing one of Vader's key shortcomings in such scenarios. While the positive class performance remains lower, this targeted correction effectively enhances overall robustness.

4.1.5 Word2Vec

```
Model Accuracy: 0.9629

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.98      0.96      5183
           1       0.98      0.95      0.96      6033

    accuracy                           0.96     11216
   macro avg       0.96      0.96      0.96     11216
weighted avg       0.96      0.96      0.96     11216
```

The Word2Vec SkipGram model, combined with a neural network classifier, achieves an impressive accuracy of 96.29%, showing balanced precision and recall across both classes. This setup effectively captures semantic relationships between words in the reviews, enhancing performance. However, the model still struggles with negations, conjunctions, and contextual dependencies—as word embeddings

are averaged, sentence-level nuances can be lost, especially in cases involving sarcasm or mixed sentiments.

4.1.6 Word2Vec with Conjunction

```
Model Accuracy: 0.9650

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96      5183
           1       0.97      0.97      0.97      6033

    accuracy                           0.97     11216
   macro avg       0.96      0.96      0.96     11216
weighted avg       0.97      0.97      0.97     11216
```

The SkipGram-based model, enhanced with conjunction handling and classified using a neural network, achieves a strong accuracy of 96.50%. By splitting sentences at conjunctions and assigning weighted sentiment scores (e.g., 0.7 and 0.3), it captures the sentiment dynamics more effectively—especially in mixed or compound sentences. This approach helps mitigate misclassifications caused by conflicting phrases, improving over traditional Word2Vec methods in handling sentence complexity and nuance.

4.1.7 Word2Vec with Parse Tree Method

```
Model Accuracy: 0.9461

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.95      0.94      5183
           1       0.96      0.94      0.95      6033

    accuracy                           0.95     11216
   macro avg       0.95      0.95      0.95     11216
weighted avg       0.95      0.95      0.95     11216
```

The SkipGram model with parse tree-based clause weighting achieves a solid accuracy of 94.61% using a neural network classifier. This method identifies main

and dependent clauses through syntactic parsing, assigning weights of 0.7 and 0.3 respectively to better represent the sentence's core sentiment. By structurally understanding sentence hierarchy rather than relying on surface-level conjunctions, it captures nuanced sentiment flow, especially in complex or nested sentence structures—leading to more informed and accurate sentiment predictions.

4.1.8 BERT

Test Accuracy: 0.9559

	precision	recall	f1-score	support
0	0.95	0.96	0.95	5694
1	0.96	0.95	0.96	5953
accuracy			0.96	11647
macro avg	0.96	0.96	0.96	11647
weighted avg	0.96	0.96	0.96	11647

The BERT model (bert-base-uncased) combined with logistic regression achieves a strong accuracy of 95.59%, effectively leveraging contextual embeddings to capture nuanced sentiment patterns. Its bidirectional encoding allows it to understand both left and right contexts, making it more robust than traditional vector-based models. However, it tends to struggle with extremely short reviews that lack sufficient context and very long ones where the input may exceed the token limit, leading to potential information loss and reduced interpretability.

4.1.9 BERT with Conjunctions

BERT + SVM Classification Report:

	precision	recall	f1-score	support
negative	0.96	0.95	0.95	5694
positive	0.95	0.96	0.96	5953
accuracy			0.96	11647
macro avg	0.96	0.96	0.96	11647
weighted avg	0.96	0.96	0.96	11647

BERT + SVM Accuracy: 95.59%

The BERT + SVM model achieves an impressive accuracy of 95.59%, combining the rich contextual understanding of BERT with the robust decision boundaries of a Support Vector Machine. By implementing conjunction logic to split complex sentences and assigning weighted sentiments (0.7 to the primary clause and 0.3 to the secondary), the model effectively captures nuanced sentiment shifts within single statements. This hybrid strategy enhances accuracy on reviews with mixed emotions, though the performance remains comparable to other top-performing BERT-based approaches.

4.1.10 BERT with Parse Tree Method

```
BERT Parse Tree-Based SVM Classification Report:
              precision    recall  f1-score   support

    negative       0.94      0.95      0.95      5694
    positive       0.96      0.94      0.95      5953

    accuracy                           0.95     11647
   macro avg       0.95      0.95      0.95     11647
weighted avg       0.95      0.95      0.95     11647

SVM Accuracy: 94.88%
```

The BERT + SVM model using the parse tree-based clause weighting approach achieves a strong accuracy of 94.88%, indicating its capability to handle complex sentence structures. Instead of relying on simple conjunction splits, this method leverages syntactic parsing to identify the main and dependent clauses, assigning them weights of 0.7 and 0.3 respectively. This allows the model to prioritize the dominant sentiment in more linguistically structured ways, improving interpretability and accuracy for syntactically rich sentences. Although slightly lower in performance compared to the conjunction-based variant, it offers a more grammatically informed sentiment assessment.

4.1.11 Sentence BERT

```
Sentence-BERT + SVM (RBF) Classification Report:
              precision    recall  f1-score   support

    negative       0.96      0.97      0.96      5183
    positive       0.97      0.97      0.97      6033

    accuracy                           0.97     11216
   macro avg       0.97      0.97      0.97     11216
weighted avg       0.97      0.97      0.97     11216

Accuracy: 0.9661198288159771
```

The Sentence-BERT + SVM model achieved the highest accuracy of 96.61%, outperforming other approaches in both precision and recall. By leveraging Sentence-BERT's ability to generate semantically meaningful sentence embeddings and combining it with the SVM classifier's strength in handling high-dimensional data, this method excels at capturing the overall sentiment even in contextually nuanced sentences. Its sentence-level embedding helps avoid issues faced by token-based models in handling mixed or indirect sentiment. This makes it a robust choice for sentiment analysis tasks where preserving semantic relationships is crucial.

4.1.12 Fine-Tuned BERT

The fine-tuned BERT model was trained on a balanced dataset of approximately 60,000 reviews, equally split between positive and negative sentiments. Over 3 epochs, the model demonstrated consistent improvement, with a notable reduction in training loss and increase in accuracy, highlighting BERT's strong contextual understanding when adapted to domain-specific data. Fine-tuning enabled the model to learn sentiment nuances specific to the dataset, leading to enhanced performance compared to using pre-trained embeddings alone. This approach proves particularly effective for custom datasets, providing a tailored and context-aware sentiment analysis solution.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1. INTERPRETATION OF RESULTS

The experimentation and evaluation of various sentiment analysis techniques provided a comprehensive understanding of their strengths and limitations. Traditional lexicon-based methods like VADER, while simple and interpretable, suffered from limitations such as inability to handle misspellings, missing lexicon entries, and lack of a consistent threshold, leading to misclassification especially in mixed or ambiguous sentiments. Machine learning approaches such as TF-IDF with Logistic Regression improved over VADER by addressing some of these issues but still struggled with complex sentence structures. The introduction of Word2Vec embeddings with neural network classifiers significantly improved accuracy, especially when combined with sentence structure-based weighting (conjunction and parse tree methods), achieving results as high as 96.5%. BERT-based methods outperformed earlier models due to their contextual understanding, with BERT + SVM and Sentence-BERT achieving accuracies above 95%, and fine-tuned BERT further improving performance through domain-specific training. Each successive model addressed specific weaknesses of the previous ones, demonstrating a clear progression toward better sentiment comprehension and classification accuracy.

5.2 LIMITATIONS AND CONSIDERATIONS

Despite the high performance achieved by advanced models like Sentence-BERT and Fine-Tuned BERT, several limitations were encountered during the project. Firstly, large transformer models are computationally expensive, requiring significant hardware resources and training time, especially for fine-tuning on large datasets. Secondly, while syntactic parsing and clause weighting improved results for complex sentence structures, they introduced additional preprocessing overhead and relied on correct dependency parsing, which itself can introduce errors. Another limitation was the handling of sarcasm, slang, or domain-specific

terminology not well represented in pre-trained embeddings. Models also occasionally failed on extremely short or overly long reviews due to insufficient or overwhelming context. Lastly, while the dataset was balanced and cleaned, real-world data is often messier, and performance might degrade without similar preprocessing. These factors must be carefully considered when deploying sentiment analysis systems in real-world applications.

5.3 FUTURE WORK

Future work can focus on improving sentiment classification performance by exploring more advanced transformer models such as RoBERTa or DeBERTa, which have shown superior results in many NLP tasks due to their enhanced pretraining strategies and deeper contextual understanding. These models can be fine-tuned on larger and more diverse review datasets to improve generalization and robustness across different product categories and writing styles. Additionally, the emergence of large language models (LLMs) like GPT and LLaMA opens up opportunities for few-shot or zero-shot sentiment classification. These models, when quantized or fine-tuned for efficiency, can serve as powerful tools for nuanced sentiment understanding, even in low-resource or real-time environments.

REFERENCES

1. **Term-Weighting Approaches in Automatic Text Retrieval**
   Salton, G., & Buckley, C.
   Information Processing & Management, 24(5), pp. 513–523, 1988.

2. **SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining**
   Esuli, A., & Sebastiani, F.
   Proceedings of the 5th Conference on Language Resources and Evaluation (LREC), 2006.

3. **VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text**
   Hutto, C. J., & Gilbert, E.
   Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM), 2014.

4. **Efficient Estimation of Word Representations in Vector Space**
   Mikolov, T., Chen, K., Corrado, G., & Dean, J.
   Proceedings of the International Conference on Learning Representations (ICLR), 2013.

5. **Attention is All You Need**

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.

Advances in Neural Information Processing Systems, 30, pp. 5998–6008, 2017.

6. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 4171–4186.

7. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**

Reimers, N., & Gurevych, I.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3982–3992.