TU WIEN Informatics

Sentimentanalyse im E-Commerce

Entwickeln eines Modells mithilfe der Verarbeitung natürlicher Sprache

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Harilla Pango, BSc
Matrikelnummer 11937902

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Allan Hanbury, PhD

Wien, 12. Februar 2025

Harilla Pango	Allan Hanbury

TU WIEN Informatics

Sentiment Analysis in e-commerce

Developing a model using Natural Language Processing

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Harilla Pango, BSc
Registration Number 11937902

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Allan Hanbury, PhD

Vienna, February 12, 2025

Harilla Pango	Allan Hanbury

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Harilla Pango, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 12. Februar 2025

Harilla Pango

Danksagung

Zuallererst möchte ich meinem Betreuer, Univ. Prof. Dr. Allan Hanbury, danken, der mich durch diese Arbeit führen konnte, indem er mir auf Anfrage unschätzbare Hilfe, Unterstützung und Feedback gab. Sein umfangreiches Wissen und seine Expertise haben es mir leichter gemacht, diese Arbeit zu vollenden, auf die ich sehr stolz bin.

Zweitens möchte ich meiner Familie danken, sowohl meiner Mutter als auch meinem Vater, die mir das Studium an der TU Wien ermöglicht haben. Ich werde ihnen ewig zu Dank verpflichtet sein und ich hoffe, dass diese Arbeit sie stolz auf das macht, was ich erreicht habe.

Zu guter Letzt möchte ich meinen Freunden danken, die mich auf diesem Weg unterstützt und mir sehr wertvolle Einblicke und Feedback zu dieser Arbeit gegeben haben.

Acknowledgements

First and foremost, I would like to thank my supervisor, Univ. Prof. Dr. Allan Hanbury, who was able to guide me through this work by giving invaluable help, support, and feedback whenever requested by my side. His extensive knowledge and expertise made it easier for me to complete this work, which I am extremely proud of.

Secondly, I would like to thank my family, both my mother and father, who made it possible for me to study in TU Wien. I will be in their debt forever and I hope that this work will make them proud for what I have achieved.

Last but not least, I would like to thank my friends, who supported my through this journey and provided very valuable insight and feedback on this work.

Kurzfassung

Diese Arbeit stellt einen Ansatz vor, der zur Verbesserung der Sentimentanalyse (SA) im E-Commerce durch fortschrittliche Techniken der natürlichen Sprachverarbeitung beiträgt. Wir verwenden zwei Basismodelle und ein drittes benutzerdefiniertes Modell: eine regelbasierte VADER-Implementierung, ein vorab trainiertes RoBERTa-Modell und ein benutzerdefiniertes, fein abgestimmtes RoBERTa-Modell, das speziell auf E-Commerce-Bewertungen zugeschnitten ist.

Die Studie befasst sich mit zwei primären Forschungsfragen: dem effizienten Umgang mit komplexen Sprachmustern in E-Commerce-Datensätzen und der Erweiterung eines multimodalen Sentimentanalysemodells. Unser fein abgestimmtes Modell zeigt eine gute Leistung beim Erkennen nuancierter Ausdrücke wie Sarkasmus und gemischter Gefühle. Die mehrsprachige Erweiterung, die mit XLM-RoBERTa implementiert wird, zeigt vielversprechende Ergebnisse in Italienisch, Deutsch, Spanisch und Französisch, allerdings mit bemerkenswerten Leistungsunterschieden.

Wir leisten einen Beitrag zu diesem Bereich, indem wir eine detaillierte Feinabstimmungsmethode entwickeln, insbesondere zur Verbesserung der Erkennung neutraler Stimmungen – ein traditionell herausfordernder Aspekt von SA. Unser Bewertungsrahmen bietet umfassende Vergleiche zwischen verschiedenen Ansätzen. Dies hilft beim Aufbau neuer Benchmarks für mehrsprachige SA im E-Commerce-Bereich. Unsere Arbeit zeigt nicht nur bedeutende Fortschritte bei einsprachiger SA, sondern erwähnt auch die verbleibenden Herausforderungen beim sprachübergreifenden Verständnis von Stimmungen, die klare Richtungen für zukünftige Forschung und Beiträge zur Optimierung mehrsprachiger Modelle und zur Integration kultureller Kontexte vorgeben.

Abstract

This thesis presents an approach that contributes to the improvement of Sentiment Analysis (SA) in e-commerce through advanced Natural Language Processing techniques. We use two baseline models and a third custom model: a rule-based VADER implementation, a pre-trained RoBERTa model, and a custom fine-tuned RoBERTa model that is specifically tailored for e-commerce reviews.

The study addresses two primary research questions: the efficient handling of complex language patterns in e-commerce datasets and the extension of a multimodal Sentiment Analysis model. Our fine-tuned model shows good performance in detecting nuanced expressions like sarcasm and mixed sentiments. The multilingual extension, which is implemented using XLM-RoBERTa, shows promising results across Italian, German, Spanish, and French, though with notable performance variations.

We contribute to the field by establishing a detailed fine-tuning methodology, particularly improving neutral sentiment detection – a traditionally challenging aspect of SA. Our evaluation framework provides comprehensive comparisons between different approaches. This helps in building new benchmarks for multilingual SA in the e-commerce domain. While demonstrating significant advances in monolingual SA, our thesis also mentions the remaining challenges in cross-lingual sentiment understanding, which provide clear directions for future research and contribution in multilingual model optimization and cultural context integration.

Contents

CHAPTER 1

Introduction

In this chapter, we set up the background and the roadmap that the thesis will follow, so that readers have a foundation for understanding the underlying problem, motivation, and solution that we provide through this work. Table 1.1 shows a list of all the abbreviations that are commonly used throughout the thesis:

Table 1.1: List of abbreviations

Abbr.	**Description**	**Abbr.**	**Description**
SA	Sentiment Analysis	NLP	Natural Language Processing
AI	Artificial Intelligence	NLTK	Natural Language Tool Kit
MSA	Multilingual Sentiment Analysis	BoW	Bag of Words
VADER	Valence Aware Dictionary and sEntiment Reasoner	BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach	XLM-RoBERTa	Cross-lingual Language Model RoBERTa
MSA	Multilingual Sentiment Analysis	ROC	Receiver Operating Characteristic
AUC	Area Under the Curve	ML	Machine Learning

1.1 Problem Statement & Motivation

Being part of Natural Language Processing, Sentiment Analysis is also known as opinion mining [Liu22]. Its main purpose is to find and classify emotions in text data, therefore facilitating the evaluation of emotional tones, opinions, and attitudes expressed in written or spoken language.

Sentiment Analysis offers several benefits. It helps companies to make data-driven decisions, improve products, and implement successful marketing plans by means of insightful analysis of public opinions and client comments [AAK+22].

Traditional SA approaches struggle with the nature of customer reviews, which frequently consist of nuanced expressions, such as sarcasm and mixed sentiments [KKKS22]. While existing SA tools provide a basic functionality, they still fail to capture the subtle contextual nuances that are specific to e-commerce reviews [KKKS22].

First, e-commerce reviews often contain product-specific terminology and context that readers may not understand. Second, the global nature of e-commerce platforms requires multilingual sentiment analysis models that are capable of undertaking much more complex tasks of international languages, as businesses increasingly operate across linguistic and cultural boundaries.

These challenges are further reinforced by the massive global scale of e-commerce operations and customer reach. Major platforms process millions of reviews daily, requiring SA solutions that are both accurate and computationally efficient. Misreading consumer mood may have significant financial consequences that influence anything from customer service policies to product development.

1.2 Research Question 1

How efficiently can the proposed fine-tuned Sentiment Analysis model handle complex language and context in e-commerce data, such as sarcasm, to improve the accuracy of sentiment identification?

RQ1 explores the capability of language models to understand and correctly classify complex sentiment expressions in e-commerce reviews. It investigates whether fine-tuning transformer-based models can significantly improve sentiment analysis accuracy compared to traditional approaches. RQ1 addresses several critical aspects:

- Complex Language Understanding

  - Detection and interpretation of sarcastic expressions
  - Handling of mixed sentiment within single reviews
- Accuracy Improvement Targets
  - Enhancement over traditional sentiment analysis approaches
  - Reduction in misclassification of neutral sentiments
- Model Efficiency Considerations & Balance between accuracy and performance

1.3 Research Question 2

How effectively can the extended model support multi-lingual and cross-lingual sentiment analysis, enabling businesses to analyze customer sentiments across different languages?

RQ2 addresses the growing need for multilingual sentiment analysis capabilities in global e-commerce operations. It examines the effectiveness of extending sentiment analysis models across linguistic boundaries while maintaining accuracy and reliability. RQ2 explores multiple dimensions of multilingual sentiment analysis:

- Cross-lingual Capabilities
  - Transfer learning between languages
  - Maintenance of sentiment accuracy across languages
- Technical Challenges
  - Preservation of sentiment during translation
  - Processing of different grammatical structures

These questions are strongly correlated to each other, with RQ1 establishing and building a strong foundation in monolingual Sentiment Analysis, and RQ2 that is addressed by further extending the fine-tuned model in RQ1. This link allows for a systematic investigation of both the fundamental challenges in Sentiment Analysis and their extension to multilingual concepts.

1.4 Methodology

We start by analyzing and conducting a literature review on Sentiment Analysis. This will help in establishing a good understanding of the current State-of-the-Art technologies and methodologies that are being used by researchers and practitioners of this field. We identify the current limitations that SA models are prone to and find gaps or limitations that these models possess. By doing so, we leverage from these findings on

their current weaknesses and can further identify the necessary steps to fine-tune these models and achieve better performance results. Our research methodology combines theoretical foundations with practical implementation, structured in the following components:

Theoretical Framework:

- Literature Review Focus Areas
  - Sentiment Analysis Evolution
  - Natural Language Processing Foundations
  - Machine Learning Algorithms in Sentiment Analysis
  - Deep Learning in Sentiment Analysis
- State-of-the-Art Review
  - VADER: Rule-based sentiment analysis
  - BERT/RoBERTa: Transformer architectures
  - XLM-RoBERTa: Multilingual models
  - Recent advancements in cross-lingual sentiment analysis

Technical Framework:

- Three distinct approaches
  - Rule-based sentiment analysis using VADER & Transformer-based analysis using pre-trained RoBERTa
  - Custom fine-tuned RoBERTa model optimized for e-commerce
  - Development of a multilingual extension using XLM-RoBERTa
- Data Collection, Preprocessing & Analysis
  - Comprehensive preprocessing of e-commerce review data
  - Implementation of robust tokenization strategies
  - Development of sentiment classification pipelines
  - Integration of multiple evaluation metrics
- Model Architecture
  - Fine-tuning of transformer architecture for domain specificity
  - Implementation of attention mechanisms for context understanding
  - Development of multilingual processing capabilities

  - Integration of cross-lingual transfer learning techniques

- Experimental Design
  - Systematic comparison of model performances
  - Cross-validation across different languages

- Evaluation Framework
  - Accuracy and Macro F1-score metrics
  - ROC-AUC analysis for each sentiment category
  - Performance analysis across multiple languages

This methodology combines theoretical research with practical implementation, ensuring both academic rigor and real-world applicability. The approach is designed to address our research questions while advancing the current state-of-the-art in e-commerce sentiment analysis.

CHAPTER 2

Literature Review

In this chapter we conduct research on Sentiment Analysis, Natural Language Processing, and Machine Learning Algorithms, which will help in developing a fine-tuned model that can enhance text classification on different datasets. For e-commerce companies to have better decision-making and improve their products & services, they have to extract insightful information from large datasets, which in this case will be product reviews from their customers.

2.1 Sentiment Analysis

Sentiment Analysis is an NLP technique that is used to identify and determine the mood expressed or the emotional tone in textual data. One can classify them as negative, neutral, or positive by examining word and phrases through available SA models that are currently offered to users [JTM+24]. This concept is also known as Opinion Mining, which analyzes people's opinions, feelings, and reviews towards specific products/services. By leveraging from the usage of SA, both small and large scale businesses can take necessary and/or precautionary steps based on feedback from their customers or potential new clients, like building recommendation systems that target these clients based on their feedback or reviews on previous purchases.
Different business sectors or industries that could leverage from using SA include - but are not limited to: airlines - to analyze customer reviews and flight experiences [FR22], hotels - to better understand and build a clear picture on what their customers/tenants enjoyed or disliked the most [MCK21], healthcare sector - to predict the likelihood of emerging mental illnesses or clinical outcomes [DR23], stock market - to predict and potentially determine the price/value or trend of a stock or cryptocurrency [MAK+23].

By classifying customer feedback and reviews as positive, neutral, or negative, businesses can gather invaluable information, which would help them not only keep up with current

market trends, but also create them. Additionally, these companies can better manage their online reputation by trying to understand consumer sentiment and biases. In a world that is quickly evolving and is becoming more digitally connected, public opinion and important topics are also essential and need to be classified by SA, since they help in promoting more intelligent and responsive decision-making [PCZ+22].
With the application and approaches of Deep learning and Machine Learning in sentiment classification, SA has seen significant improvements in recent years [RAA+22]. Because of these deep neural networks and conventional machine learning algorithms, which we will also briefly describe in Chapter 3, the precision and scalability of SA systems has increased. Therefore, it is essential that we analyze the state-of-the-art in the next chapter, where we cover the recent experiments and developments in SA, its application domains, datasets, pre-processing techniques, ML, and DL [Hus18].

Limitations of SA in its current stage include: lack of ability to analyze sarcasm/satire, using slang or inappropriate grammar style, limited data available mostly in English, which brings inaccurate results that cannot be interpreted to derive and extract meaningful conclusions [JTM+24]. We explore a lot of these limitations in detail later on in Chapter 3 of our thesis.
A key component of SA is context-based sentiment, or also known as Context Understanding, which takes into account the context or subtle meaning behind certain words, which affect the sentiment conveyed in text [JTM+24]. Emotions may be very context-based or dependent and get influenced by sarcasm, ambiguity, irony, negation etc. For example:

"I must say, I am thoroughly impressed with the XYZ Vacuum Cleaner 3000. It's so lightweight and portable that it barely even picks up dirt from my carpet. I love how the battery only lasts for a measly 10 minutes before needing to be recharged for hours. It's perfect for those who enjoy spending more time charging their vacuum than actually using it."

The intended attitude is negative, even though there are positive words in the sentence, like "love", "perfect" etc. At the same time, same words can have different meanings, for example: "bank" (river bank vs. a financial institution). This requires word disambiguation. Ambiguity makes thing more complicated and difficult for SA systems, because of the intended meaning of statements with various implications, just like the word we gave as an example above.
[ZZCC20] claim that it is challenging to make a distinction between implicit and explicit sentiments in text/sentences, since explicit emotions can be easily identified by sentiment words. On the other hand however, implicit emotions are much more difficult to deal with in the absence of such terms. This is where creating sentiment lexicons come in hand, but those require a large amount of labelled data and resources, which make textual emotion even more challenging.

2.1.1 Sentiment Analysis Levels

a. Document level analysis
This level is used primarily to give polarity to a whole document. It can be used if someone wants to give a general analysis and classify a page or chapter of a book as positive/negative/neutral [BJE15]. Irrelevant statements that need subjective or objective categorization are among the challenges or drawbacks of document-level analysis. By taking advantage of features like opinion words and term frequency, both supervised and unsupervised learning methods can be used [BAB21]. Some of the supervised methods include:

- **Linear regression**: predicts numerical values by fitting a line to data points
- **Random forests**: helpful for classification and regression tasks
- **K-Nearest Neighbors (KNN)**: makes use of the average value of the K nearest training examples in order to make predictions
- **Naïve Bayes**: a probabilistic classifier based on Bayes' theorem

Some of the unsupervised methods include:

- **K-Means clustering**: splits data into k distinct groups by iteratively assigning points to the nearest cluster center
- **Hierarchical clustering**: creates a tree-like hierarchy of clusters
- **Association Rule Learning**: examines relationships between variables in large-scale datasets

b. Sentence level analysis
Each sentence of a document, article, or review is individually specifically analyzed. Particularly useful if this document has a plethora of sentiments or range of feelings towards a specific topic/event [YC14]. This level uses far more trained data and processing resources, but still uses the same techniques/methodologies like the ones in document-level. This level is critical and specifically helpful when dealing with ambiguous statements and/or conditional sentences [FE19].

c. Phrase level analysis
Phrase level revolves around analyzing the sentiment of a given group of words, or phrase, rather than each individual word. This approach could be very useful in our research, since it can help analyze specific opinions on a sarcastic level, by taking into account the context in which the words or phrases are used. Furthermore, phrase level analysis offers a more complex understanding of the emotional tone connected to specific words. This is useful in scenarios where emotions or feelings might change within a sentence.

Learning-based, lexicon-based, and hybrid techniques are among the several algorithms utilized to solve sentiment classification issues [MAY+24].
Strengths - better suited and used for applications with nuanced analysis, such as product reviews or customer feedback.
Drawbacks - slower and more complex to implement.

d. Aspect level analysis
[CXX+21] define Aspect Based Sentiment Analysis (ABSA) as the task of categorizing the sentiment polarity of a certain sentence element. The goal of ABSA is to identify those sentiment elements in a text, which could consist of one or multiple elements with associated dependency connections. In early ABSA research, when there was a need to extract sentiment information about a certain element from the context of the text, Neural Models were used as the primary tool for extraction [LSG+22].
Additionally, ABSA seeks to determine whether sentiment elements - such as aspect category or sentiment polarity - are relevant to the text [ZLD+22].

2.2 Natural Language Processing

The goal of Natural Language Processing (NLP), a branch of artificial intelligence and linguistics, is to enable computers to comprehend words or assertions written in human languages. It was created to make the user's job easier and to fulfill their desire to speak to the computer in natural language [KKKS22]. Figure 2.1 shows NLP's broad classification and how it is split into two main components [KKKS23]:

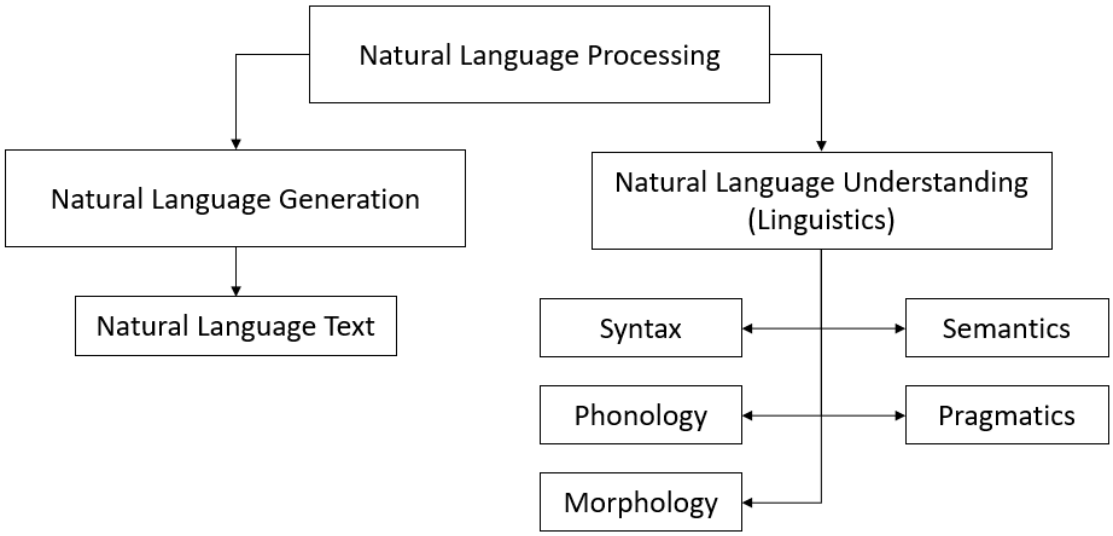

Figure 2.1: Main Components of Natural Language Processing

While other professionals have also shown interest including linguists, psychologists, and philosophers etc., most of the work in NLP in the current literature is carried out by computer scientists [KKKS23]. NLP is associated with several ideas and methods addressing the issue of natural language of communication with the machines [KKKS22]. Some of the researched tasks of NLP are:

- Automatic Summarization - generates a logical synopsis
- Discourse Analysis - the task of determining the related text's discourse structure
- Part of Speech - tells a sentence's length and finds each word's component of speech
- Optical Character Recognition - translated printed and handwritten text into machine-readable format to enable automated text recognition
- Machine Translation - the task of automated text translating from one language into another
- Co-reference Resolution - a group of text that identifies every word referring to the same thing

2.2.1 Natural Language Generation (NLG)

The process of creating meaningful words and sentences from an internal representation is known as Natural Language Generation, or NLG[KKKS22]. NLG consists of four stages:

- Goal identification
- Planning for goal achievement
- Accessible communication channels
- Textual realization of the plans

2.2.2 Natural Language Understanding (NLU)

In customer service applications, NLU is utilized to comprehend issues that consumers have expressed orally or in writing. Through the extraction of concepts, entities, emotions, keywords, and other elements, it makes it possible for machines to comprehend and interpret natural language[KKKS22]. Its most important terminologies are:

- **Syntax** - a statement demonstrating the structural dependencies between words is the result of this level. Another name for it is parsing, which identifies sentences that have greater significance than individual words. At the syntactic level, words are organized into phrases, which are subsequently grouped into clauses, and finally, phrases are joined to sentences. Since eliminating stopwords alters the sentence's meaning, this level keeps them in place.

- **Phonology** - [Las84] stated that phonology is "the study of sound pertaining to the system of language". The subdiscipline of linguistics is therefore concerned with the behavior and organization of sounds. A crucial component of phonology is the semantic use of sound to encode meaning in any human language. The term "phonology" comes from ancient Greek language, and is made of two words "phono" - which means voice or sound - and "logos" - which means speech or word.

- **Morphology** - the study of morphology examines how words are constructed from morphemes, which are smaller, meaningful units. Morphemes can be categorized into two major groups. Stems, which are the word's fundamental, significant units and its root. On the other hand, affixes give words extra meanings and grammatical purposes [KKKS22]. Humans may break down any unknown word into morphemes to comprehend its meaning, but the interpretation of morphemes remains consistent across all words.

- **Semantics** - From a semantic standpoint, the most important chore from a sentence is figuring out its proper meaning. To grasp a sentence, humans rely on their language knowledge and the concepts it contains. The semantic level assesses words based on their dictionary interpretation—that which results from the phrasing context. Thus, the remaining part of the phrase is under close review to identify the appropriate interpretation; only then will one be able to ascertain the actual meaning of the statement [Fel99].

- **Pragmatics** - this level emphasizes on the knowledge or material derived from beyond what the document contains. It examines the less direct spoken phrases or sentences. Analysis of the context helps one to develop text representation. Pragmatic ambiguity arises when a phrase is not specific and the context offers no particular knowledge about that sentence [Wal13]. Additionally, the background of a text could include the references to other lines of the same document. While pragmatic analysis concentrates on the inferred meaning that the readers perceive depending on their previous knowledge, semantic analysis stresses on the literal meaning of the words.

2.3 Machine Learning Algorithms

2.3.1 K-Nearest Neighbors

The Nearest Neighbor Classification idea is straightforward: instances are arranged based on the class of their closest neighbors. Using k-nearest neighbors helps to define the class as it is often useful to take into account several neighbors. Since induction is postponed until run time, it is considered as a lazy learning method. Since it is based just on training examples, it is also known as Example-Based Classification or Case-Based Classification. The training examples must be in memory during runtime [CD07].

2.3.2 Naïve Bayes

Probabilistic models, based on the simplifying presumption of solid independence [R+01], are derived from the Bayes theorem. Naïve Bayes is essentially based on the computation of category probabilities for a given text document by use of joint probability assessment of words and categories. Many times employed in supervised classification are Bayesian network classifiers [FGG97]. Originally noted in text retrieval, it is still a fundamental method in text classification and categorization. Two advantages of the Bayesian classification method are the combination of historical data knowledge with observed data and provision of helpful learning algorithms. Bayes theorem for probability has a basic formula as follows:

$$P(y|x) = \frac{P(x|y)}{P(x)} \tag{2.1}$$

2.3.3 Random forest

A supervised Machine Learning technique for classification and regression problems is Random Forest. It is a group of algorithms that use decision trees as individual predictors. A large number of decision trees are trained and their predictions are then combined, in order to generate an outcome [Bre01]. The algorithm is therefore trained with multiple decision trees, each fed with slightly different data samples. Random forest also includes Random Subspace, randomizing outputs and boosting it. Numerous decision trees are created by learning techniques during the training phase. Additionally, the output is aggregated from each tree to find class mode and it can then be used for both regression and classification.

2.3.4 Support Vector Machine

Support Vector machines (SVMs), a class of machine learning algorithm, are enabled by performing optimal data transformations that define boundaries between data points based on predefined classes, labels, or outputs, so addressing difficult classification and regression tasks. Speech and image identification, healthcare, signal processing applications, etc. are only a few of the industries where SVMs find the most usage [HDO+98]. The SVM formula is given below:

$$f(x) = \sum_{x_j \in S} y_j K(x_j, x) + b \tag{2.2}$$

where:
xi - training patterns,
yi - class labels,
S - Set of support vectors

2.4 Deep Learning Algorithms

2.4.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are characterized by a loop topology, which is made up of interconnected neurons. RNNs are efficient at applications that require sequential data processing, because of the fact that they have innate memory capability that allows them to handle input sequences in a more efficient and skillful manner than standard feedforward neural networks [MJ+01]. An RNN can carry out a consistent operation for each element in a sequence thanks to the concept of "memory". What this means is that each output that the RNN is producing will be impacted by all previous calculations. Therefore, it stores history information in a manner that is similar to how people remember and recall knowledge from the past.

2.4.2 Convolutional Neural Networks

Artificial neural networks called Convolutional Neural Networks, or CNNs, are frequently used for computer vision and image recognition tasks. The three main layers of CNNs are:

- **Convolutional layers** - by computing the scalar product between the weight of the neurons and he input region, convolutional layers determine the output of neurons connected to specific regions of the input.
- **Fully connected layers** - they seek to generate class scores for categorization/-classification based on the activations.
- **Pooling layers** - by performing spatial downsampling, the pooling layer reduces the activation's parameter count.

CNNs are inspired by the human visual system and they can independently learn and identify certain patters in images without prior knowledge [O'S15].
What follows below is a formula for the computation of a convolutional operation's outcome:

$$C_{i,j} = \sigma(b_j^l + \sum_{m=1}^{M} W_{j,m}^{l,j} X_{i+m-1}^{l-1}) \tag{2.3}$$

where:
l - the layer index,
σ - the activation function,
b - a feature map's bias term,
M - kernel size,
W - a feature map's weight

2.4.3 Long Short-Term Memory networks

Long Short-Term Memory Networks are advanced recurrent neural networks. They provide a well-constructed structure by creating "gates" in their fundamental unit, which are also known as "cells". These gates can prevent gradient bursting and/or disappearing in conventional RNNs and capture both short-term and long-term memory along the time steps. These gates are named as: "forget gate", "input gate", and "output gate" [Hoc97].

2.4.4 Gated Recurrent Units

An alternative to LSTMs, is the Gated Recurrent Unit (GRU), which is an example of an RNN architecture that deals specifically with sequential data processing [HS97]. The internal design of GRU and LSTM is where they diverge the most. Although the goal of both architectures is to capture long-term dependencies, LSTM contains three - forget, input, and output). On the other hand, GRU has two - update and reset - gates, which makes it an even simpler structure that the previous. By adding gating methods, it solves the vanishing gradient issue with conventional RNNs. In some situations, GRU is more effective than LSTM. It frequently uses fewer parameters, which results in less computational costs. In other applications though, LSTM is able to catch complex relationships more effectively[HS97].

CHAPTER 3

State-of-the-Art

In this chapter, we provide an understanding of the current state-of-the-art of Sentiment Analysis in recent years, with an emphasis on the latest papers and state-of-the-art. We then dive into numerous applications of SA, their most commonly used datasets, pre-processing techniques or methodologies, while also providing information about existing pre-trained models. Furthermore included in this chapter is an analysis of current developments and researcher contributions that provide an understanding of the state-of-the-art experimental results impacting the field of the study.

3.1 Related Work

Often known as opinion mining, SA is a fundamental part of NLP. SA's powers also include more complex techniques that allow one to detect certain emotions, intents, or subtle nuances of sentiment, including happiness, wrath, sarcasm, or context-specific sentiments noted in fields including product evaluations [CDBF17].
Results of a detailed review of contemporary opinion mining literature can be found in the work of [SLZ+21] as well as in a published overview by [LCT21].
For machine learning, the authors in [HF17] make a brief presentation and overview of the algorithms used in social media analysis. Additionally, for social media platforms that fully utilize advanced searching algorithms, there is an examination on how they can be applied, which is conducted by [BAB21].
Authors of [SGJ+17] and [YV20] have both published papers on sentiment classification, while also covering and investigating the area of fraudulent reviews, for example: a sole proprietor or entrepreneur decides to open an e-commerce business, selling different products/services and after some time, they gain a significant amount of the market share. Therefore, competitors would try to re-gain advantage, be it in ethical or unethical ways, such as: sending bots and spamming product review pages with negative comments and reviews, which would have a huge impact in driving the sales of that particular

store/business down.
On the other hand, in order to determine how effective internet reviews are, authors [YCL+19] and [LZ12] have already conducted a research. There have also been a various number of surveys, where the authors [YCL+19] and [BKBH21] have made their suggestions after identifying the main problems of sentiment analysis. In the paper of [PMS17], we can find a study and discussion of the topic that ranges from 2000 to 2015. It provides a framework of the processing of unstructured data computationally, with the main focus of extracting views based on their moods and feelings.

3.2 Real-world Applications

Sentiment Analysis can be used across multiple platforms or industries, as long as there is user-generated content or comments/opinions.
As mentioned in [ZO18], Sentiment Analysis has already been applied previously different domains, which range from stock markets and healthcare, to hotels and airlines etc. in order to better understand the customer's or tenant's likes and dislikes.
In the paper [SLZ+21], they cover how to represent knowledge in opinions, how to extract text from reviews/opinions with uncertainty and how to categorize them. Additionally, they also show the results of a detailed review of contemporary opinion mining literature.
A method for adaptable aspect-based lexicons for sentiment classification is proposed by [MAK20]. The writers outlined two methods for creating two dynamic lexicons to help categorize emotions according to their aspects: genetic and statistically based approach algorithms. According to [NKU21], a dynamic lexicon offers more accurate grading for ideas connected to context and may be updated automatically. Based from several dictionaries, they selected a number of lexicons, so that organization of all aspects of the reviews was done easier.
Analyzing tweets and their sentiments, that are related to different topics or global events is crucial, just liked mentioned and already applied by authors in [AAAA19].
Regarding the health care sector, which has lately seen a surge in SA applications [DR23], there have been multiple papers that cover customer opinion analysis [RKK20] [PPC20] [CD21] and also customer satisfaction analysis [BAP+20] [MWW+18].
In the finance industry, apart from stock markets, cryptocurrencies have always gained a lot of attention from people all over the world. Being very volatile and heavily dependent on "hype" or speculation, users' or traders' opinions matter mostly more than anything. This is where we see SA being used in [VGEVA19], where it determines the trends of cryptocurrencies and stock markets [ZO18] based on emotions and market sentiment.

We will list some of the key real-world applications that take advantage of SA and leverage on it, with the primary goal of extracting as much information as possible from their user datasets. In the next chapter, we will see how our dataset is specifically extracted and processed to fit an e-commerce business. However, our fine-tuned model is designed to work with other large datasets too, which means that users will be allowed to input their own dataset from different domains, like movie review datasets, twitter comments,

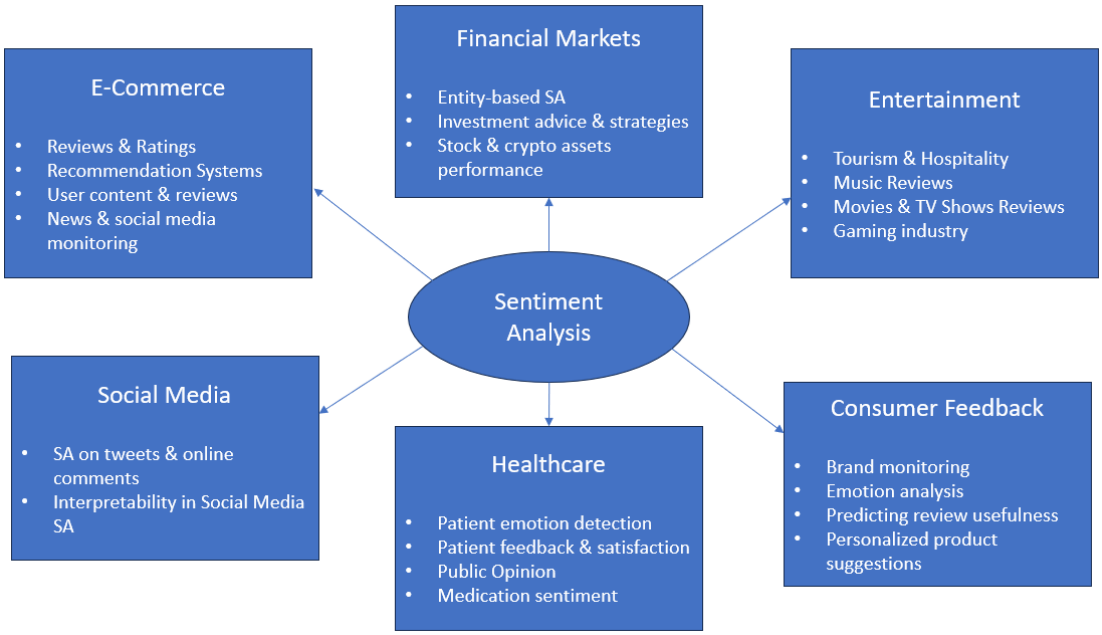

Figure 3.1: Application of SA in different domains [SAR23]

financial stock comments and speculations etc. Therefore, we also mention some of the most significant domains that SA covers in Figure 3.1:

3.2.1 Product reviews in e-commerce

The e-commerce industry has been steadily growing for the past decade [RR21], showing no signs of stopping, especially after the events and restrictions of COVID-19, where most of the world was on lockdown and people were "forced" or had to resort to other ways of gaining access to physical goods/services, thus shifting their attention to e-commerce businesses. Giants that are dominating the world market, like Amazon, Alibaba, eBay, and Walmart are trying to find ways of improving their already existing recommender systems. One way to do this is through SA and NLP. Their attention has therefore shifted towards using the best state-of-the-art models, or further improve them to achieve better understanding on their consumers' sentiment. Having already collected a large amount of data from their customers, these companies are able to leverage on this and make better decisions for their business model or develop new marketing plans [HLYJ19]. In[BDRS20], the authors came to the conclusion that SA may be used to identify customer behaviors and hazards and increased customer satisfaction.

As the e-commerce world continues to expand, so do the numbers of products sold and

reviews coming in from customers all over the world, so this is where SA comes into power, which can give very good insight to businesses and help customers choose better products [Par03]. Using sentiment analysis, a company may ascertain the opinions of its customers on a recently launched product or by analyzing reviews and comments. One option is to choose keywords related to a particular product characteristic/feature (food, service, cleanliness), and then train a sentiment analysis framework [MMC15] to recognize and evaluate just the relevant data on an aspect-level based approach [SF15].

[EABH23] customized e-commerce chatbot using Sentiment Analysis, therefore addressing gaps in current research & offering customized user experiences. [KS23] designed a strong hybrid technique for SA to give companies a tool to evaluate social attitude surrounding their goods or services, thereby overcoming uncertainty and inconsistencies in language. [KKO+23] enhanced the correctness & accuracy for recommendation systems by introducing the Hybrid Recommendation Model (HRM). For accurate sentiment multi-classification in reviews, a directed weighted model was employed by [ZZZW20], providing effective classification with constrained threshold criteria. A flexible framework for researching this sentiment analysis method was provided too. [YSY22] provided e-commerce insights and useful recommendations for businesses, by using SA based on public sentiment to analyze Twitter data from Lazada and Shopee. Furthermore, [JSV21] demonstrated useful uses of SA for real-time product suggestions and insights gleaned from customer reviews, and suggested using PySpark and the flexible distributed dataset (RDD) for SA.

Recent research focuses on machine learning techniques for multilingual SA [SPMM20], Arabic SA [OADAH21], implicit aspect extraction for SA [GR19], social media SA [DK19], machine learning techniques for SA [ASKP20], and lexicon-based and Bidirectional Encoder Representations from Transformers (BERT)-based SA in Italian [CPE22]. There is a lack of thorough research on SA in e-commerce, including a summary of methodologies, overview of data platforms, and probable future approaches [HAM23]. We explore later on in chapter 4 how we contribute to the solution of this problem, where our fine-tuned model takes care of multi-lingual SA tasks using the base model of BERT and improving it further and achieve good results on multiple languages - spanish, italian, german, and french.

There are several ways of how SA can be applied in e-commerce, like ML methods (SVMs and Naïve Bayes), which call for less data but more human participation [HAM23]. On the other hand, DL methods such as BERT and LSTM depend mostly on large data rather than feature engineering [DWT+20]. We see how we apply these methods and techniques in chapter 4, where we take a large dataset of more than 500K reviews from Amazon products and then explain in detail what this dataset contains.

3.2.2 Stock market

In the financial industry, day traders and market analysts are always on the look for new ways to predict future price actions regarding a particular stock/cryptocurrency. Although there are already existing methods to do that, like Technical Analysis (TA) [HLZZ24] - a financial analysis that uses patterns in market data to identify trends and make predictions, Sentiment Analysis comes into play. It leverages the use of all the news and "hype" or speculation surrounding a stock trend. In this case, data is abundant and mostly free, since it can be gathered by different news articles, blogs, forums, Twitter etc. By analyzing all these opinions, text can then be polarized and categorized accordingly. In [XCW18] we see methods being used to determine whether the trend points to be increasing or decreasing. In financial markets, positive news usually lead to upward movement [OV11], whereas negative news often cause panic and can lead to downward movement, especially in the case of cryptocurrencies.
There are a few studies and researches that use sentiment analysis to the field of blockchain technology, as in the work of [KDS20]. The impact of news sentiment on the volatility, volume, and average returns of traditional currencies and cryptocurrencies such as Bitcoin, Ethereum, XRP, and ripple was examined by [RHZ20].

3.2.3 Entertainment, Hotels & Restaurants

Another sector that makes use of SA is the Entertainment industry, which includes TV shows, movies, and short films. SA can be used in this case, as it is designed to understand sentiments of viewers' comments and therefore it helps in making a better choice on what content they want to get shown in their "Recommended For You" page - as Netflix often does with their software [KYR19].
Sentence level SA deals with providing a better understanding of the overall and average sentiment of reviews from users/viewers [LH09]. Additionally, [JYPS21] have also discussed categorizing consumer decisions as either positive or negative based on their provided online reviews [JSP21].
On the other hand, hotels and restaurants have also implemented SA over the last few years. Through customer reviews, they can create better recommendation systems and improve client experience [ZXW19]. Service providers can largely profit from aspect-based SA - a method of text analysis separating the text data and defining its sentiment depending on its features [KJ23], since they can extract the aspect that receives the most negative feedback and improve on it [SL20] [ASQAA+18].

3.2.4 Healthcare & medical domain

Even though not as widely used as in other domains, SA continues to expand in this industry [DR23]. Domain experts are researching different methods to find more uses of SA and NLP [EXT+21]. If one wants to analyze new updates in the medical field, they can do so by collecting large amounts of data from articles, news, Twitter [CP21], blogs, and most importantly surveys.

Furthermore, applications of SA can help service providers of this domain collect and better understand patient epidemics, drug reactions, and diseases. Tweets have previously been used to analyze patient experiences as an add-on for public health in [CJJ+18]. Using Twitter's Streaming API, the authors [CJJ+18] gathered around 5 million tweets on breast cancer over the course of a year. Following pre-processing, the tweets were classified using a CNN model and a typical LR classifier.

3.2.5 News & Social Media

There is already a wide range of applications in English language, where short texts like tweets or movie reviews are analyzed [KJ20, BKBH21]. On the other hand, Pereira [Per21]. and [KBK+21] discuss how there currently is a gap for non-English short texts, compared to standard English.
In order to fill this gap, [ZAZC18] offers a benchmark evaluation of 28 academic and commercial Twitter Sentiment Analysis (TSA) systems using five different Twitter datasets, as well as a thorough overview of the present status of TSA systems and approaches. The study's objectives are to evaluate these systems' effectiveness, pinpoint their advantages and disadvantages, and draw attention to the major issues and developments in the TSA industry.
[ZAZC18] focuses on benchmark evaluation, which covers a wide range of TSA systems from general-purpose to domain-specific techniques. These systems use a variety of strategies, including deep learning architectures, machine learning algorithms, and lexicon-based approaches. The benchmark results reveal that domain-specific systems consistently outperform their general-purpose counterparts, with an average accuracy of 67% compared to 56% [ZAZC18].
Furthermore, some of the best-performing systems, like BPEF and Webis, use ensemble techniques that integrate many classifiers, and others, like NRC, use a wide range of lexical resources to improve their sentiment analysis performance [ZAZC18].
Even with the improvements in TSA procedures, a number of significant obstacles still exist. For fine-grained sentiment interpretation, precise aspect/phrase-level sentiment analysis and target recognition are also essential [ZAZC18].

3.3 Large Language Models (LLM)

In NLP, some of the most essential LLMs include GPT, DialogueLLM and InstructionERC models. These models offer a wide range of language-related tasks, such as text production, summarization, machine translation, question answering, and chatbot interactions. Additionally, they use sophisticated transformer topologies with more than 100 billion parameters. Most notably, OpenAI's GPT-3 is a model that has proven to be remarkably adept in capturing subtleties of language patterns, comprehending text, while at the same time, producing coherent text. These LLMs leverage the potential of DL and ML on a variety of textual data. Therefore, they have changed the SA research landscape.

The contribution of LLMs to the advancement of SA research is covered in detail in one survey by [ZWL18]. This survey emphasizes how LLMs can perform better than conventional methods, particularly when working with big datasets. [CLC20] conducted another survey. It explores how LLMs affect SA tasks and it also mentions the fact that DL techniques have replaced conventional methods. The author puts emphasis on how LLMs helped improve sentiment prediction accuracy and capture semantic link.

3.3.1 Bidirectional encoder representations from transformers (BERT)

Pre-trained generic LLMs, including BERT, have shown extraordinary performance across a range of NLP applications. It is built on Transformer architecture, which means that it uses attention techniques in its encoder/decoder components [AM20]. By incorporating bidirectional contextual embeddings using a transformer-based architecture, which makes use of masked language modeling during pre-training, BERT transformed NLP, because it was able to capture deep semantic relationships in text. Crucial for understanding sentiment and semantic subtleties in diverse language contexts, and in order to process input tokens simultaneously, the model uses multi-head self-attention mechanisms with position embeddings and segment embeddings, while keeping awareness of their relative positions. Since BERT is not meant for sequence-to- sequence activities, it is imperative to underline that it mostly uses the transformer encoder and excludes a decoder network.

Strengths:

- Capable of transfer learning
- Robust treatment of polysemy, which means words that have many meanings
- Strong performance in several fields following fine-tuning (as we see later in Chapter 5)
- Pre-trained on large-scale data
- Can be fine-tuned for specific tasks with limited datasets

Drawbacks:

- Needs a high amount of GPU memory and resources (which we take care of by using Google Colab's T4 GPU)
- Bad handling of terms from outside of formal language/vocabulary
- Computationally costly
- Difficulties with domain-specific lingo without particular fine-tuning

3.4 Simple rule-based Model

3.4.1 Valence Aware Dictionary and sEntiment Reasoner (VADER)

VADER is a simple rule-based model for general SA. When the data being analyzed is unlabeled, VADER may identify the polarity of the sentiment and categorize it as either positive or negative. This is one drawback of using this model, as it is strictly limited only to that, compared to other models, where they can classify sentiments as negative, neutral, or positive. The VADER algorithm is allowed to learn from the labelled training data. Predicting a movie review's star rating based on a comment or critic's evaluation is a classic example.
On the other hand, we chose VADER because it also has its own strengths as a model. Even emojis like ":-(", or extended punctuation (? vs. ???) are some of the many examples of non-conventional text that VADER can recognize and comprehend. Another capability of VADER is that it doesn't need a lot of preprocessing to function. Requirements like tokenization and stemming are not necessary, in contrast to many other supervised NLP techniques.

Strengths:

- Results are interpretable because of its lexicon-based methodology
- Functions well without the need for GPU resources from scratch
- Rule-based methodology that eliminates the need for fine-tuning or training data
- Especially tuned for casual writing & social media text
- Good in using language criteria to identify sentiment intensity

Drawbacks:

- Restricted knowledge of sarcasm
- Unlike ML models, cannot learn from fresh data
- Performance mostly depends on the completeness and quality of its data/vocabulary
- Depends mostly on established vocabulary
- Needs more in-depth contextual knowledge

3.5 Methodological Challenges

Dealing with ambiguous situations and irony—that is, sarcastic statements about an item that are meant to communicate a negative attitude but are frequently misread by typical sentiment analysis algorithms—are two major issues in sentiment analysis [CHPR+19] [MHK14]. Due to its cultural distinctiveness and the challenge of robots comprehending distinct and sometimes highly nuanced cultural allusions, sarcasm detection is still an open problem. According to [PHC+18], multimodal sentiment analysis may detect sarcastic remarks more frequently if it incorporates voice and facial emotions.
For societal reasons unrelated to their underlying opinions, people may convey sentiments in order to define and express their identities or to conform to a certain issue norm [Mcl23]. What this means is that people may voice specific opinions in order to fit a given social group or identity, therefore defining themselves [BMS11]. For instance, someone may say strong opinions about environmental concerns in order to be viewed as environmentally sensitive. Even if individuals privately feel otherwise, people may voice ideas that line up with what they believe to be the acceptable or prevailing viewpoint in their social circle or group [YQG+19].
Similar datasets with the same size as those used for emotion identification are needed for sentiment analysis of human-to-machine and human-to-human interactions. Despite having a lower audiovisual capture quality, [MEKCP14] show how webcams may be utilized to gather a lot of emotional reactions, including sentiment.
Word emotion information is not taken into account by word embedding techniques such as word2vec and GloVe, which translate words into vectors [WAV22]. Below we list some of the key problems in SA and NLP, and briefly explain the terms:

Sarcasm
Generally speaking, sarcasm may seem as a praising thought or opinion towards something or someone, but in reality, it is quite the opposite. Sarcasm is mostly used by people that are disappointed and when their expectations haven't been met. Being a satirical remark, it doesn't only revolve around creativity in usage of words, but other factors also affect it, such as situational context, tone, and background information. It is used often when people want to criticize and express implicit information, with the sole purpose of mocking something or hurting someone's emotions. This is why NLP still faces lots of unique challenges and obstacles in properly analyzing sarcastic text/comments. Because of these challenges, NLP has become very interesting lately as a research subject, due to its usefulness in enhancing social media SA [ENSN20].

Computational Cost
When using big datasets and various languages, fine-tuning transformer models like RoBERTa and XLM-RoBERTa are prone to major GPU memory and processing capability.

Informal style of writing
Pre-trained models such as RoBERTa may underperform on informal language including emoticons, acronyms, and colloquialisms, hence misclassifying emotion.

Grammatical errors
Ungrammatical content can compromise transformer models' contextual comprehension, therefore influencing their capacity to properly capture sentiment.

Adaptation of Languages
There are hundreds of languages and different variations of the same language spoken worldwide today, but resources that are available out there are mostly in English. But even being available, there are varieties of English that are spoken in other regions, such as, American, British, Indian etc. Although the base remains the same, there can be noticeable differences and changes, like pronunciation, literacy rate, prominence etc. For example, the term for flip-flops or slippers is "thong", however in UK that very same word has a meaning for undergarments. Additionally, different spellings of the same word, like "color" and "colour", have different regional spellings even if they both imply the same thing. This will result in duplication and might have an impact on the model's computational cost and most importantly, accuracy.

Phrases that contain intensifiers and degree adverbs
Adverbs like barely, mildly, and faintly are employed to quantify the feelings. Let's take 2 example reviews:

Review 1 = "The food is barely good" and,
Review 2 = "The food is really good"

While R2 is seen as extremely positive, R1 is thought to be neutral or somewhat positive. The degree of positivity and the word "good" are determined by the adjectives "barely" and "really". In the same way, intensifiers measure the phrases' emotional content. To boost the token's positive or negative aspect, intensifiers such as extremely, too, are employed. For example, it is thought that "too good" is more favourable than "good". Rather than distinguishing between two phrases with opposing polarities, intensifiers and degree adverbs make it difficult to aggregate the sentiment values and compare two sentences with the same feeling.
When it comes to technical sentiment difficulties, however, the most commonly applied method is the N-gram methodology, which is based on words and expressions [WWH09]. Furthermore, the lexicon-based approach is the least employed strategy. When addressing the specific sentiment difficulties, the theoretical challenges use a range of strategies to improve performance [HCF12]. According to [TBT+11], the theoretical kind extensively uses lexicon-based techniques and PoS tagging.

CHAPTER 4

Model Creation

In this chapter, we cover the practical part of our thesis and describe how we build our fine-tuned model. We start with the Data collection & Pre-processing. Since we want to build a model that is specifically tailored to e-commerce tasks, we use a large dataset of over 500K reviews. Even though the fine-tuned model can process other datasets (social media comments, IMDB movie reviews etc.), we expect to achieve the best results mostly with e-commerce based data, since all the steps and methodologies that were taken to build this model were particularly designed for e-commerce SA.
We then show all of the pre-processing techniques or methods that we use in order to achieve better results with our model, compared to the VADER and BERT models. The Python code that we wrote in Google Colab can be found here: Notebook Link.
After building and fine-tuning our model, we get all the results and compare them: VADER vs. RoBERTa vs. Fine-tuned RoBERTa model. We evaluate them using metrics, like: Accuracy, F1-score, AUC and ROC, and Confusion Rate. This indicates if the model can perform well for analyzing the sentiment and sarcasm of e-commerce based reviews.

Therefore, we have divided our work into the following blocks/steps:

Step 1 - Data Collection & Preprocessing
Step 2 - Baseline SA with VADER & RoBERTa
Step 3 - Fine-tuning the RoBERTa Model
Step 4 - Extending our model for Sarcasm Detection
Step 5 - Model Training & Evaluation
Step 6 - Extending our model for multilingual SA

4.1 Data Collection & Preprocessing

4.1.1 E-commerce platform APIs and Data dumps

A lot of the e-commerce companies like Amazon, Alibaba, eBay and Walmart give access to APIs (Application Programming Interfaces). However, a lot of these APIs require users to pay a small fee in exchange, or register and send an application, which usually takes some time for it to get approved by the respective company. Examples:

1. Amazon Product Data: https://registry.opendata.aws/amazon-reviews/
2. eBay Product Data: https://go.developer.ebay.com/api-documentation

4.1.2 Web scraping using Python

Some APIs are not available or too expensive to access, so we also consider switching to web scraping. By developing or even using already available scraping tools, we can crawl all necessary web pages and automatically extract all the data. However, this comes at a cost, since it is very important to note that we have to comply and respect website terms of service and check if there are any applicable policies/legal restrictions that apply to web scraping. Examples of scraping tools:

1. Beautiful Soup (Python library for web scraping): https://www.crummy.com/software/BeautifulSoup/
2. Scrapy (Python web scraping framework): https://scrapy.org/

4.1.3 Publicly available datasets

There are also many publicly available datasets that researchers or organizations can use, like:

1. Amazon Product Data: https://nijianmo.github.io/amazon/index.html
2. Yelp Dataset: https://www.yelp.com/dataset
3. Multi-Domain Sentiment Dataset: https://www.cs.jhu.edu/ mdredze/datasets/sentiment/
4. Kaggle (one of the most famous open-source platforms): https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

4.1.4 Our proposed dataset

As mentioned above, we have to comply and agree to the Terms & Services of certain websites that crawl for data. The websites that we tried to scrape already mentioned that they don't allow scraping for data, therefore we skip this possibility and instead switch to an ethical way of gathering our data. We make use of a large Amazon dataset that contains more than 500K product reviews, ranging from 1 to 5 star reviews. The reason for choosing this particular dataset is because of its good rating distribution (as we see below from the Exploratory Data Analysis of this dataset). Reviews include product and user information, ratings, helpfulness scores, full plain text reviews, and also an 'isSarcastic' label, which is denoted as either a '1' or a '0' - 1 meaning the review is sarcastic and 0 meaning the review isn't sarcastic. This dataset has also been used in other research papers [YSS22] [AVG23] [Sha19] and can also be found across different repositories on GitHub and YouTube tutorials.

4.1.5 Data Preprocessing

Data pre-processing is the process of text preparation for categorization or classification by means of cleaning, since the text may often have plenty of noise and useless elements like HTML tags or scripts. On language level, these sentences may have little effect on its overall meaning, therefore necessary cleaning is required to make them usable when analyzing the sentiment behind them.
Online text cleaning, white space removal, expanding abbreviation, stemming etc. are some of the steps take in order to pre-process our dataset. All of these processes are called transformations [FHM08]. The idea of having the data pre-processed is simple: to lower the noise in the text, which should assist in enhancing the performance of the classifier and speed up the classification process, thereby supporting real time analysis. We list below the pre-processing techniques that we use in our fine-tuned model:

- **Tokenization** - a basic first step in NLP and SA, where this technique separates text into individual words or tokens [Ali12]. We use tokenization in both the base model and the multilingual extension, which is implemented through XLMRobertaTokenizer and RobertaTokenizer.
- **Data cleaning** - the process of spotting and fixing mistakes or inaccuracies in datasets. It improves their quality and dependability for use in analysis [RD+00]. We use data cleaning through the tokenizer's preprocessing capabilities. This way we handle basic text cleaning as part of the RoBERTa tokenization process.
- **Normalization** - this is the process of standardizing measurements and enabling accurate comparisons by means of numerical data adjustment to a range between 0 and 1 [Pat15]. Normalization is applied in our model through layer normalization layers and used in both sentiment classification and sarcasm detection heads.

- **Feature extraction** - reduces dimensionality by means of a new representation that captures necessary information for analysis while modifying data [GE06]. We implement feature extraction through the transformer architecture's self-attention mechanisms and it is used in sarcasm detection through the dual-head architecture.

4.1.6 Exploratory Data Analysis (EDA) of the dataset

ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I hav bouç seve the Vitali cann d...
B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Prod arriv label Juml Salte Pear
B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This conf that beer arou fe...
B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If yo looki for th secr ingre i...
B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Grea taffy grea price Ther was wid..

Figure 4.1: Dataset EDA

From Figure 4.1, we focus only on what is important and what plays an impactful role in our model, therefore the most important ones will be the text and the 'isSarcastic' label. We graphically show the Review Length Distribution, the Review Length by Rating, and the Review Scores Distribution for 10000 reviews of our dataset in Figures 4.2, 4.3, and 4.4, respectively. The reason for using only 10000 reviews is because using all 500K reviews would require large GPU memory and processing time for a dataset of that scale, surpassing the current limitations of conventional research settings and cloud computing platforms like Google Colab, which usually impose memory constraints of 12-16GB GPU RAM. While reducing the dataset size, we made sure to keep the original distribution of review scores and lengths from the 500K dataset.

Figure 4.2 provides a clear picture of how these reviews are distributed. In detail, we can see that the histogram shows a positive skew (right-skewed) distribution of the character length, where most of the reviews are in the 0-1000 range. This distribution also shows a

long tail going beyond 4000 characters. Long reviews are rare here, since we can see that the frequency drops significantly after 1000 characters.

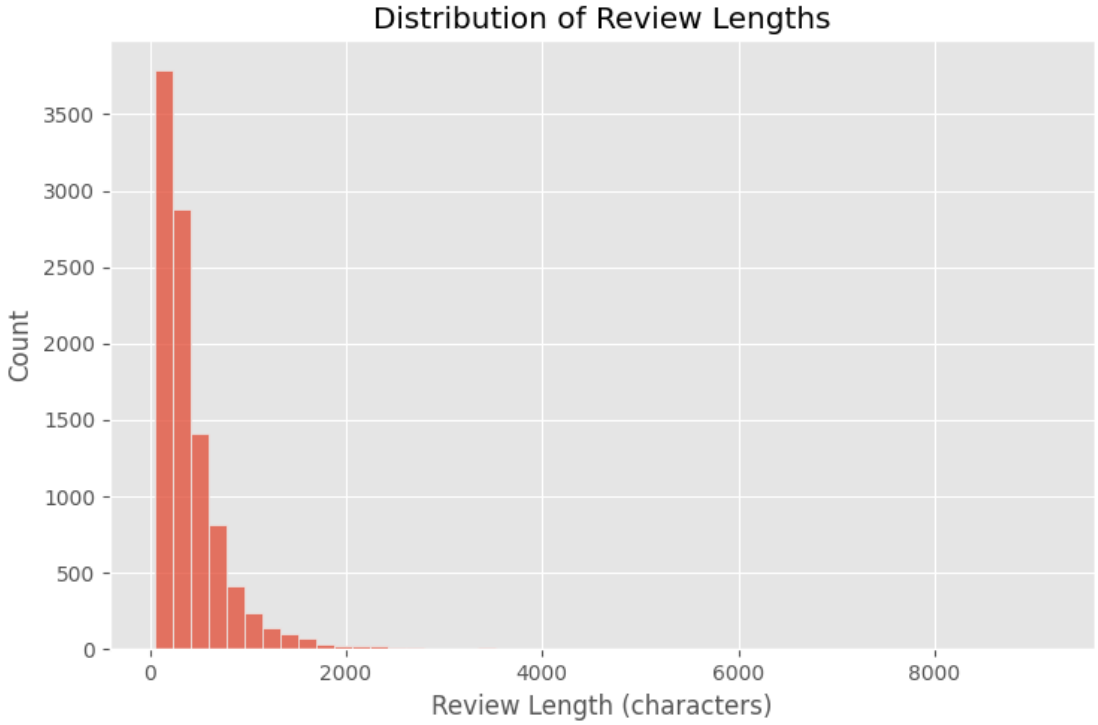

Figure 4.2: Visualization of Distributions

Additionally, the box plot in the Figure 4.3 shows the relationship between review length and rating scores. All rating categories have similar median lengths, which is indicated by the horizontal line in the box. Maximum outliers reach around 8000-9000 characters. Considering the size of the whole dataset, this figure does a good job at showing a consistency in size across ratings, suggesting similar variability in review lengths, regardless of rating.

Lastly, Figure 4.4 shows a bar chart of the distribution of review scores, where the frequency distribution of ratings is on a 1-5 scale. This shows a strong positive skew toward 5-star ratings. Here we can see a clear "J-shaped" distribution, which is common in e-commerce reviews, but at the same time, our data shows a class imbalance, with 5-star reviews being 4 times more frequent than any other rating.

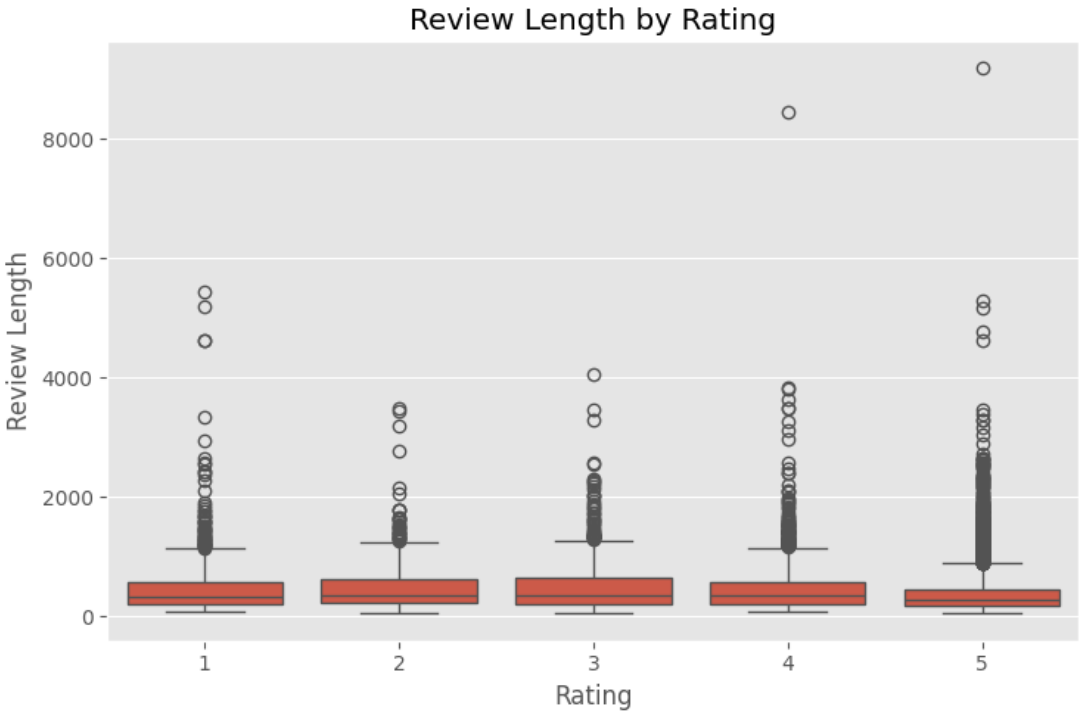

Figure 4.3: Review Length & Scores

4.2 Baseline SA with VADER & RoBERTa

The first phase of our SA model makes use of two different approaches: the rule-based VADER and the transformer-based RoBERTa model.

4.2.1 VADER Implementation

The VADER sentiment analyzer served as our primary rule-based baseline, applying its specialized dictionary and empirically determined rules for sentiment intensity scoring. VADER's lexicon is ideal for e-commerce since it includes domain-agnostic sentiment indicators such as punctuation emphasis and capitalization. The analyzer gives us four metrics for each review: positive, negative, neutral, and a compound score ranging from -1 (very negative) to +1 (highly positive).
In our implementation, we used VADER's compound score as the primary metric for sentiment classification - since it provides appropriate sensitivity to sentiment shifts while being resistant to noise in the text, with thresholds of > 0.05 for positive sentiment, $<=$ -0.05 for negative sentiment, and values in between for neutral classification.

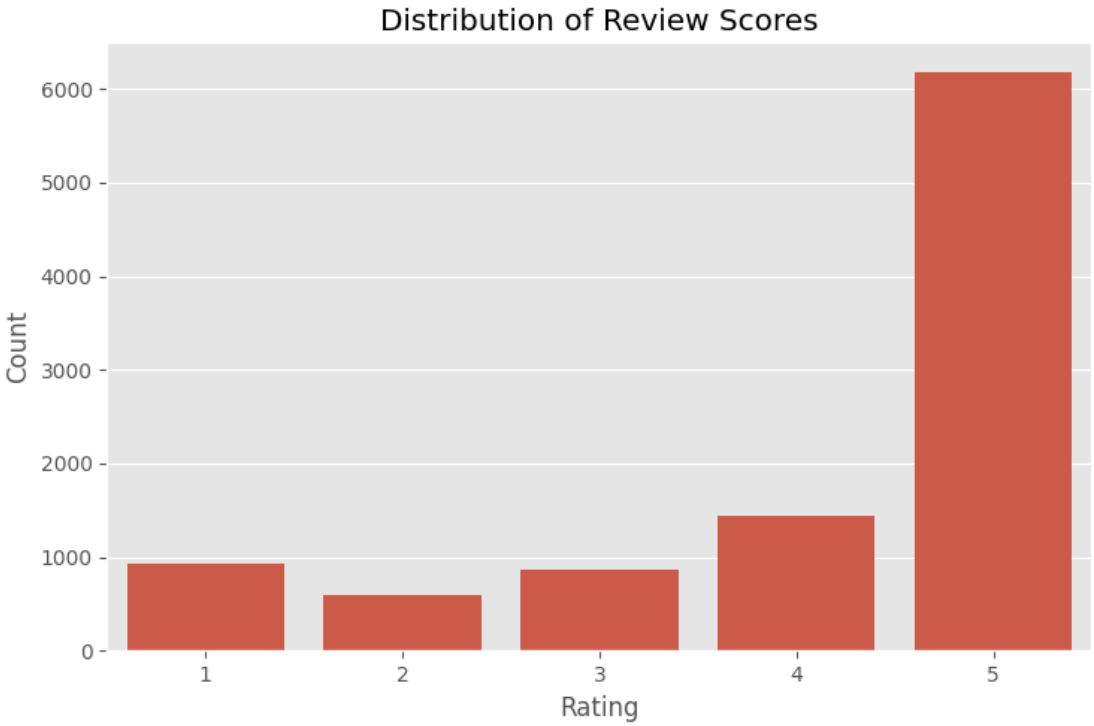

Figure 4.4: Distribution of Review Scores

4.2.2 RoBERTa Implementation

The second baseline uses the RoBERTa model, specifically the 'cardiffnlp/twitter-roberta-base-sentiment' pre-trained variation, which has been effective in social media sentiment analysis. The transformer-based design uses bidirectional contextual representations and attention mechanisms to identify complicated linguistic patterns and long-range dependencies in the review text.

Our RoBERTa implementation includes:

1. Tokenization and Encoding: RoBERTa's byte-level BPE tokenizer identifies out-of-vocabulary terms and preserves subword information for e-commerce terminology.

2. Attention Mechanism: We use RoBERTa's multi-head self-attention architecture to capture contextual connections and semantic dependencies in review text.

3. Sentiment Classification: The model generates probability distributions for three sentiment classes (negative, neutral, and positive) using a softmax activation function, allowing for discrete classification and confidence score.

4.3 Fine-Tuning the RoBERTa Model

4.3.1 Custom Dataset Architecture

We used a specialized SentimentDataset class using PyTorch's Dataset interface to enable effective fine-tuning. This approach includes a number of design choices: in order to balance computational efficiency with the preservation of semantic value in e-commerce reviews, the dataset implementation uses dynamic tokenization with a maximum sequence length of 128 tokens. While preserving model performance, this sequence length optimization takes into consideration the verbose character of product reviews.

4.3.2 Model Architecture Improvement

Our fine-tuning approach includes several architecture improvements to the base RoBERTa model:

- Contextual Feature Extraction: On top of RoBERTa's output embeddings, we added a bidirectional LSTM layer with 256 hidden units. The reason for this architectural choice is the necessity to capture long-range relationships and sequential patterns unique to e-commerce reviews—where sentiment frequently changes throughout the text. The bidirectional nature guarantees that the final representation contains both forward and backward contextual information.

- Advanced Regularization: A complex regularization technique combining dropout (0.2) and layer normalization is used. By successfully avoiding overfitting and preserving the model's capacity to identify domain-specific patterns, this dual strategy was especially selected to solve the difficulties of fine-tuning big language models using domain-specific data.

4.3.3 Training Optimization Strategy

The fine-tuning process incorporates different optimization techniques:

- Learning Rate Scheduling: We used the following parameters to create a linear learning rate scheduler with warmup: initial learning rate: 3e-5 and warmup steps: 10% of total training steps

- Loss Function Engineering: A hybrid loss function that combines a learning component with cross-entropy loss for classification is used in the training phase. By taking into account both the continuous nature of sentiment intensity and the categorical character of sentiment categorization, this method allows the model to develop more robust representations.

4.3.4 Memory Management & Computational Optimization

For effective resource use, we decided to optimize our model by:

- Gradient Accumulation: We used a batch size of 32 for gradient accumulation, which preserves the advantages of higher batch sizes while enabling efficient training on constrained computational resources. This method offers more consistent gradient updates, which is advantageous for the fine-tuning process.
- PyTorch's automated mixed precision training is used in the training pipeline, which dynamically switches between float32 and float16 calculations. This optimization preserves numerical stability while consuming less memory, which is especially important for RoBERTa's attention processes.

Additionally, in order to avoid overfitting, we included an early stopping mechanism with a three-epoch patience that monitors validation loss. Maintaining model generalization capabilities while adjusting to domain-specific (e-commerce) patterns has been successful with this method.

4.4 Extending our model for Sarcasm Detection

4.4.1 Architectural Enhancement for Sarcasm Detection

Our approach introduces a specific Sarcasm Aware Sentiment Classifier, which extends the existing architecture. This improved model includes a number of elements that are used to capture both explicit and implicit signs of sarcastic speech. Among the essential architectural elements are:

- Contextual Feature Extraction Layer:
  - Bidirectional LSTM with 256 hidden units
  - Specialized for capturing long-range contextual dependencies
  - Enhanced ability to identify sentiment inversions and contradictions
- Dual-Head Architecture:
  - Primary sentiment classification head
  - Dedicated sarcasm detection head
  - Integrated feature fusion mechanism

4.4.2 Sarcasm Detection Methodology

To find possible examples of sarcastic expression, the sarcasm recognition model uses multiple detection techniques. Several important aspects of sarcastic language in e-commerce evaluations are the focus of our implementation:

- Contextual Contradiction Analysis: The model makes use of a contradiction detection method that examines differences between the sentiment of individual sentences and the context of the entire review. Finding instances when positive language is used to express negative sentiment—a prevalent trend in e-commerce reviews—is especially important.
- Marker Recognition System: To identify linguistic and punctuation markers often linked to sarcastic expression in written text, we integrated a customized system. The system takes into account:
  - Extreme sentiment expressions
  - Punctuation patterns (e.g., multiple exclamation marks, ellipses)
  - Contextual polarity shifts

4.4.3 Performance Optimization

In order to to maintain computational efficiency even on sarcasm detection, our model extension uses the following optimization techniques:

- Attention Mechanism Optimization:
  - Optimized memory utilization through gradient checkpointing
  - Specialized attention patterns for sarcasm detection
- Training Efficiency:
  - Specialized learning rate scheduling
  - Dynamic batch size adjustment
  - Gradient accumulation strategies

4.5 Model Training & Evaluation

A neural architecture that goes beyond just employing pre-trained models is at the center of our work. Our implementation adds a few essential changes for e-commerce sentiment analysis to the RoBERTa basic architecture, which was chosen for its strong contextual understanding capabilities. The model uses a hierarchical structure in which domain-specific components are added to the basic transformer layers.

Using 12 transformer blocks with 12 attention heads each and a hidden size of 768 dimensions, the architecture starts with RoBERTa's pre-trained transformer layers. However, by adding specialized components, we significantly change this starting point. Our addition of a bidirectional LSTM layer (256 hidden units) on top of RoBERTa's output embeddings is intended to capture long-range relationships in customer reviews. In e-commerce sentiment analysis, where sentiment frequently changes over the course of a review and necessitates knowledge of product-specific context, this element is very important.

4.5.1 Training Process Implementation & Evaluation

A data splitting strategy was used in our study to guarantee accurate model evaluation. The data is split as follows:

- Training Set: 60%
- Validation Set: 20%
- Test Set: 20%

With an initial batch size of 16, the training used dynamic batch processing, which automatically adjusted according to available computing resources to maximize memory consumption while preserving training stability. Using a linear learning rate scheduler with warmup was a crucial part of our training plan. The scheduler started with a linear decay phase after a warmup that accounted for 10% of the total training steps. This strategy was crucial for maintaining training stability in the early phases and guaranteeing effective convergence in all subsequent epochs. We constructed the AdamW optimizer by using a calibrated learning rate of 2e-5 and weight decay of 0.01.
We used a comprehensive performance evaluation across several parameters, as you can also see later on in Chapter 5. Throughout training, the evaluation process was continuously monitored, with validation set evaluations being carried out at 100-step intervals. Early identification of possible problems like overfitting or unstable training dynamics was made possible by this regular monitoring. In order to present a fair assessment of our model's performance, the following key evaluation metrics were used:

1. ROC-AUC scores for individual sentiment categories
2. Detailed confusion matrix analysis for error pattern identification
3. Classification accuracy and macro-averaged F1 scores for overall performance assessment

4.6 Extending our model for multilingual SA

After considering our fine-tuned model's architecture and getting optimal results (which we describe in the next chapter), we successfully answer our first research question that is mentioned in the introductory chapter. On the other hand, in order to answer our second research question, we extend our model to support multilingual SA and compare the scores and performance results next to each other. We continue our work by breaking down how our extended model works and which key features it utilizes from a technical aspect.

We investigate additional techniques used in this extended model of our base fine-tuned model. This implementation of the model significantly extends it by enabling cross-lingual SA and maintaining model efficiency.

To begin with, we utilize XLM-RoBERTa, a multilingual version of RoBERTa, as our base model, which is pre-trained on 2.5TB of data containing 100 languages. In our version of the model, we train it using 4 languages (Spanish, Italian, German, French). This extended version implements device-agnostic computation, which has automatic GPU detection and utilization, but falls back to CPU usage when necessary, and has efficient resource management.

For our scale of the thesis, we create a custom dataset out of our Amazon reviews by using synthetic data generation through translation. We achieve this with Google Translator API for automated translation, while making sure that we maintain sentiment labels across translations and create a balanced multilingual dataset. This model can handle 10 major languages, but we specifically use only 4, as mentioned above. However, this leads to a potential limitation of being able to only use these 4 specific languages.

For the training infrastructure, we utilize a specialized multilingual training pipeline with a small batch of 16 to handle varying sequence lengths, a smaller amount of epochs (3) due to increased data diversity and a conservative learning rate of 2e-5 for stable training.

For the evaluation metrics, we continue to use the metrics used in our model, which are per-language accuracy tracking, F1-score and performance visualization through heatmaps. Additionally, we make a cross-lingual performance comparison through visual performance analysis and a detailed metrics report, all of which are shown in the next chapter.

CHAPTER 5

Evaluation & Results

In order for us to provide an appropriate analysis on the evaluation & results of all 4 models (2 base models and our 2 fine-tuned models), we plot these results and analyze them. This way, we can extract this information and clearly see our improvements in comparison to the base models.

5.1 VADER vs. RoBERTa

We start by showing the performance metrics results of the 2 base models (10K records) in Figure 5.1:

Metric	VADER Model	RoBERTa Model
Accuracy	0.648	0.675
F1 Score (macro)	0.390	0.479
AUC Score - Negative	0.770	0.763
AUC Score - Neutral	0.500	0.729
AUC Score - Positive	0.765	0.790

Figure 5.1: Performance Metrics Results

For the VADER model, we see that we achieve an accuracy of 64.8% and an F1-score of 0.390, both of which are low in comparison to RoBERTa (67.5% and 0.479, respectively). When we see a value of 0.500 in the neutral sentiment from VADER, it basically means that the model is unsure of the true sentiment of the sentence and it takes a random sentiment guess on those particular sentences. A negative sentiment AUC score of 0.770 means that there is moderate discrimination capability, and a positive sentiment AUC score of 0.765 means there is a basic sentiment detection coming from this model.
On the other hand, RoBERTa achieves better results, with an accuracy of 67.5% and an F1-score of 0.479. The AUC scores also show higher results, with a negative sentiment detection score of 0.763, a good neutral class discrimination of 0.729, and a positive sentiment detection of 0.790.

In Figure 5.2, we visualize the distribution of sentiment scores for both VADER and RoBERTa models, separated by true sentiment classes (negative, neutral, and positive). The y-axis in both plots represents the density, or relative frequency, of scores at each point on the x-axis. Higher peaks indicate more frequent occurrence of those particular scores for that sentiment class. As we can see from the left graph, the score distribution of

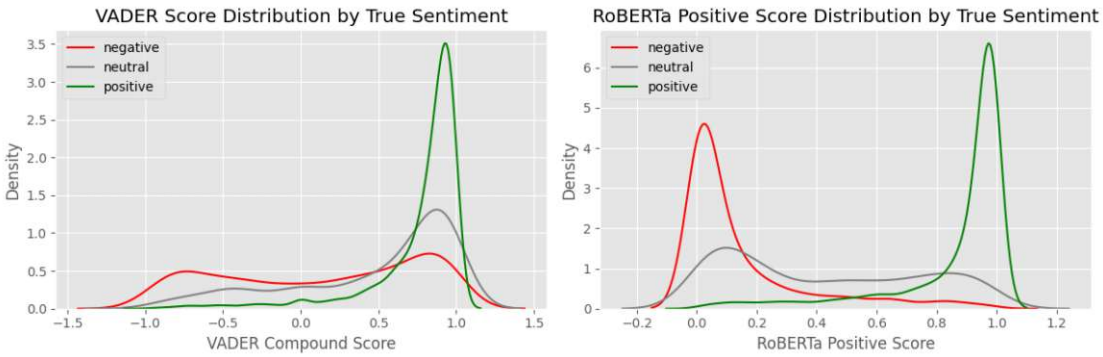

Figure 5.2: Score Distribution Graphs

VADER shows overlapping distribution between sentiment classes, with wider distribution curves that suggest lower confidence in classifications. Additionally, it is less pronounced between negative and neutral sentiments. It also shows a peak for positive sentiments around 0.9, indicating strong positive classifications. Plotted on the x-axis, we see the "VADER Compound Score" which ranges from -1 to +1. This score is derived from VADER's SentimentIntensityAnalyzer, which calculates a compound sentiment score by combining positive, negative, and neutral components.
On the other hand, the RoBERTa score distribution shows a similar distribution between sentiment classes with sharp and distinct peaks for positive (0.9) sentiments. The x-axis represents the probability/confidence score of the positive sentiment class from RoBERTa's softmax output.

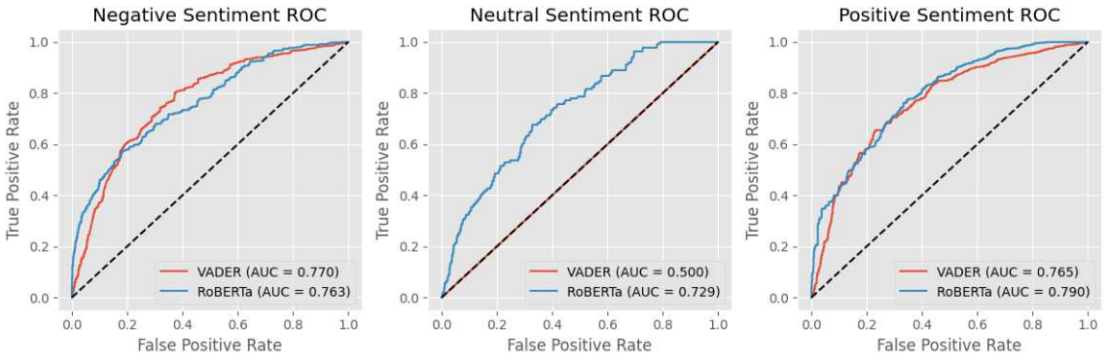

Figure 5.3: Sentiment ROC

Furthermore, from Figure 5.3, we can see that RoBERTa shows a slightly better discrimination ability. For the neutral sentiment detection, RoBERTa still maintains reasonable performance for this challenging middle category and as for the positive detection, it also shows superior performance across all operating points.

5.2 Fine-tuned Model

For our fine-tuned model, we present its performance results in this subsection, then in Section 5.4, we will compare all of the evaluation metrics to the two previous base models to see which one performs better and is more accurate across all sentiments. As

Metric	Value
Accuracy	0.952
F1 Score (macro)	0.865
AUC Score - Negative	0.994
AUC Score - Neutral	0.959
AUC Score - Positive	0.993

Figure 5.4: Performance Metrics Result - Sentiment

we can see from the performance metrics results in Figure 5.4, our fine-tuned model achieves an accuracy of 95.2% and a high macro F1-score of 0.865. A high accuracy

represents the proportion of correct predictions across all classes and indicates that our model correctly classified 95.2% of all test samples. Since the macro F1-score is close to 1.0, it indicates that the model maintains high precision and recall across all sentiment categories (positive, neutral, negative).

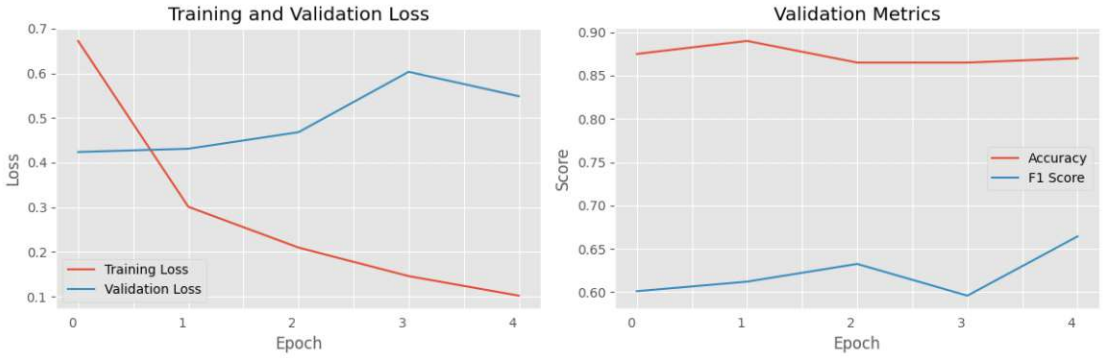

Figure 5.5: Training/Validation Loss & Validation Metrics

Starting off with the left image in Figure 5.5, it represents two key training plots, training loss and validation loss. The red line (training loss) shows consistent decrease from about 0.7 to 0.1 across epochs, whereas the blue line (validation loss) initially increases until epoch 3, then shows a slight decrease. The key observation here is that we have a strong initial learning phase from epochs 0-1, and then an optimal performance point around epoch 1-2. A slight overfitting starts after epoch 2, where validation loss increases while training loss continues to decrease.
On the other hand, the right image displays a validation metrics plot, where it shows two key performance metrics like Accuracy and F1-score, both of which gradually improve over time, as seen in the image. An important pattern here is the rapid improvement in both metrics during early epochs at 0-1. There is a stabilization of performance around epoch 2, with slight oscillations in F1-score after epoch 2. Additionally, the model maintains consistent high accuracy throughout the training.

All AUC values in Figure 5.6, across all 3 sentiments, are very close to 1.0. We have a negative sentiment of 0.994, a neutral sentiment of 0.959, and a positive sentiment of 0.993. This means that our fine-tuned model is very good at making a distinction between different sentiments with high accuracy and confidence. When the red line curves towards the upper left corner of a Receiver Operating Characteristic (ROC), it means that the model is performing in an excellent way, because the ROC curve plots the true positive rate against the false positive rate at different classification thresholds.

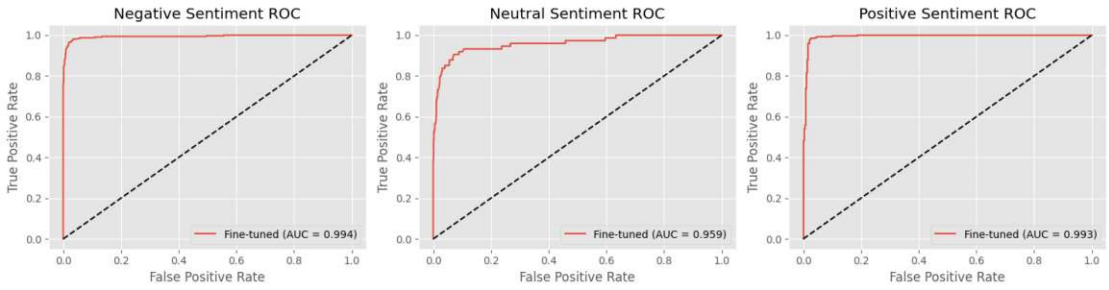

Figure 5.6: Sentiment ROC

Results - Sentiment
Accuracy Enhancement:

- 41% improvement from base RoBERTa (67.5% to 95.2%)
- 46.9% improvement from VADER (64.8% to 95.2%)

F1-Score Advancement:

- 80.6% improvement from base RoBERTa (0.479 to 0.865)
- 121.8% improvement from VADER (0.390 to 0.865)

AUC Score Improvements:

- Negative sentiment: 30.3% improvement (0.763 to 0.994)
- Neutral sentiment: 34.8% improvement (0.729 to 0.983)
- Positive sentiment: 25.8% improvement (0.790 to 0.994)

The most significant improvements were observed in:

- Neutral sentiment detection (34.8% AUC improvement)
- Overall F1-score (80.6% F1-score improvement)
- Classification accuracy (41% classification improvement)

Results - Sarcasm

Finally, Figure 5.7 shows the results for the Sarcasm Detection Model on the 10K records, where the dataset distribution was 30% sarcastic reviews and 70% non-sarcastic reviews. Results are measured on the entire test set for sarcastic reviews. Additionally, using our dual-head architecture, we evaluate how effectively the model can identify sarcastic reviews from the test set. The sarcasm detection head generates probability scores for each review, with scores above 0.5 indicating sarcastic content.

By showing both the results for sentiment and for sarcasm separately, we can understand how the presence of sarcasm affects the model's ability to determine the true sentiment of a review. This way, we analyze both the overall accuracy and the specific types of misclassifications that occurred in sarcastic versus non-sarcastic reviews (as shown in the previous results section).

Our fine-tuned RoBERTa achieves an accuracy of 0.919 on sarcastic reviews, which outperforms both VADER and base RoBERTa, which achieved a similar accuracy of 0.541. This 37.8% improvement shows that the fine-tuning process significantly enhanced our model's capacity to accurately classify sentiment in reviews that contain sarcastic elements.

The F1-score results show an even bigger improvement, where our fine-tuned RoBERTa achieves a score of 0.903, which significantly exceeds that of VADER (0.309) and base RoBERTa (0.410).

The confusion rate represents the proportion of wrong sentiment predictions on sarcastic reviews and it shows the models' robustness against sarcastic misdirection. Our fine-tuned RoBERTa achieves a low rate of 0.054, which means that it gets confused on predicting the sarcasm of a review 5.4% of the time. Compared to VADER's 0.324 and base RoBERTa's 0.243, it shows a 26.9% reduction in confusion rate from the best baseline model (RoBERTa) and it indicates superior capability in distinguishing genuine sentiment from sarcastic expression. In Figure 5.8 we can see some examples of sentiment prediction on sarcastic reviews by all 3 models.

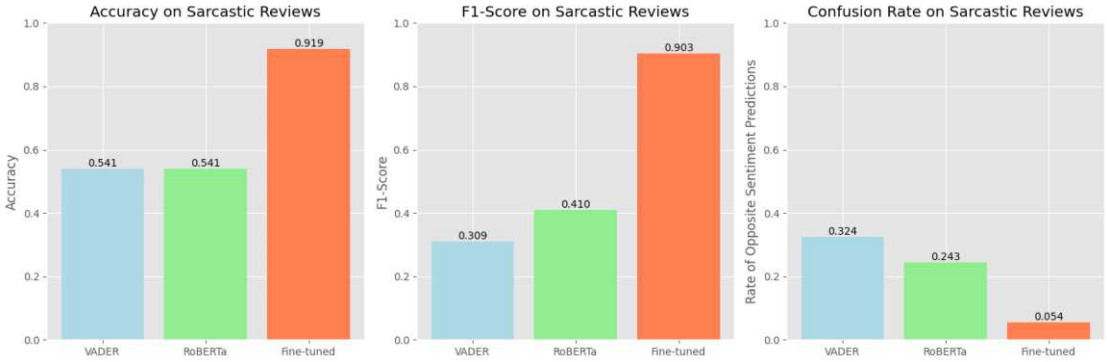

Figure 5.7: Accuracy, F1-score, Confusion Rate results on Sarcastic Reviews

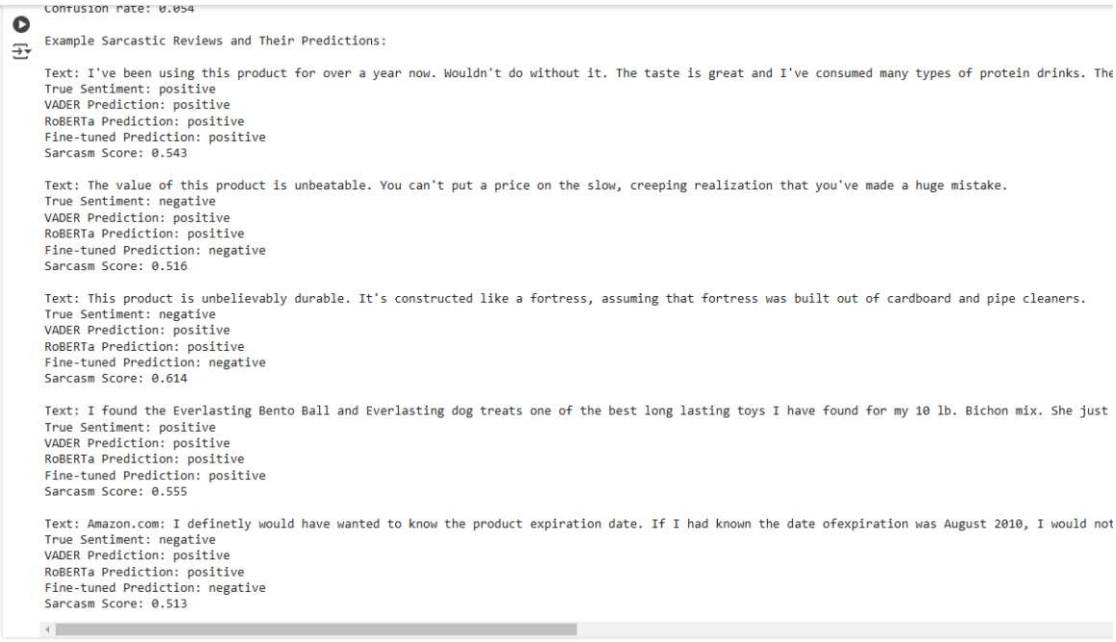

Figure 5.8: Sarcastic Reviews Examples and their Predictions

The superior performance of our fine-tuned model can be attributed to several factors:

1. Domain adaptation to e-commerce review specifics
2. Balanced training approach maintaining performance across all sentiment classes
3. Effective learning rate scheduling preventing overfitting

5.3 Multilingual Fine-tuned Model

From the results in Figure 5.9 we can see that Italian has the highest Accuracy (0.840 and F1-score of 0.498) among all languages and shows strong classification capability. German is the second best (with Accuracy of 0.832 and F1 of 0.530), which shows that has the most balanced performance between accuracy and F1. Balanced performance refers to how well the model performs across all sentiment classes (positive, negative, and neutral) rather than excelling at one class while performing poorly on others. Spanish has a good Accuracy of 0.800, but a moderate F1 score of 0.511. Lastly, French was the lowest performing language with Accuracy of 0.768 and F1 of 0.464, which shows the largest gap between them and suggests challenges in handling minority classes.

The training & loss graph in Figure 5.10 shows a steady decrease in training loss from 0.75 to 0.45, which indicates consistent learning. The validation loss is relatively stable

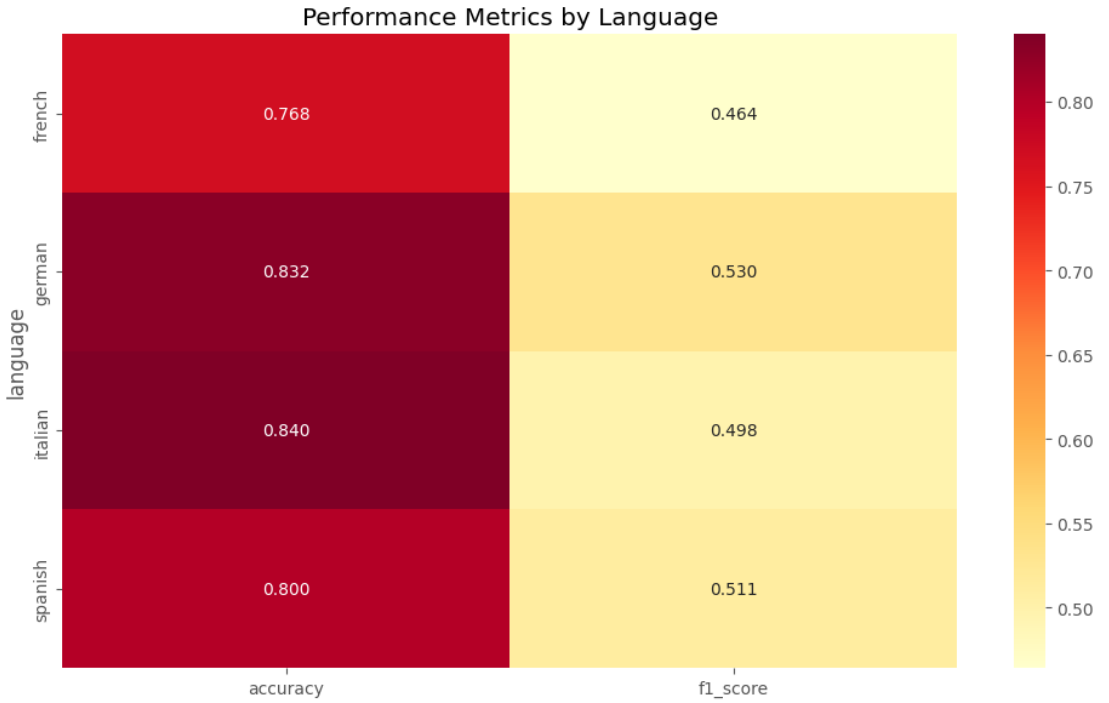

Figure 5.9: Performance Metrics by Language

at around 0.60, where minor fluctuations suggest stable cross-lingual learning. The gap with training loss indicates expected multilingual complexity.

Figure 5.11 shows us that the validation Accuracy improves from 0.75 to 0.81, with steady improvement across epochs and no plateauing is observed within training period. On the right hand side, we see that the F1-score increases from 0.30 to 0.50, showing significant improvement in balanced performance and lower than monolingual model due to increased task complexity.

Results

- Accuracy - Monolingual (0.975) vs. Multilingual (0.768-0.840) - expected drop due to cross-lingual challenge
- F1-score - Monolingual (0.942) vs Multilingual (0.464-0.530) - the larger gap indicates challenges in maintaining precision/recall across languages

Strengths of model:

Figure 5.10: Training & Validation Loss

- Consistent performance across multiple languages - the model maintains similar levels of accuracy and F1-scores across different languages without showing dramatic performance drops for any particular language
- Stable training dynamics - refers to the predictable decrease in training loss without fluctuations or sudden spikes
- Progressive improvement in both accuracy and F1-score - indicates that both accuracy and F1-scores show a steady increase over training epochs

Limitations of model:

- Significant performance drop compared to monolingual model
- Lower F1-scores indicate challenges in consistent predictions

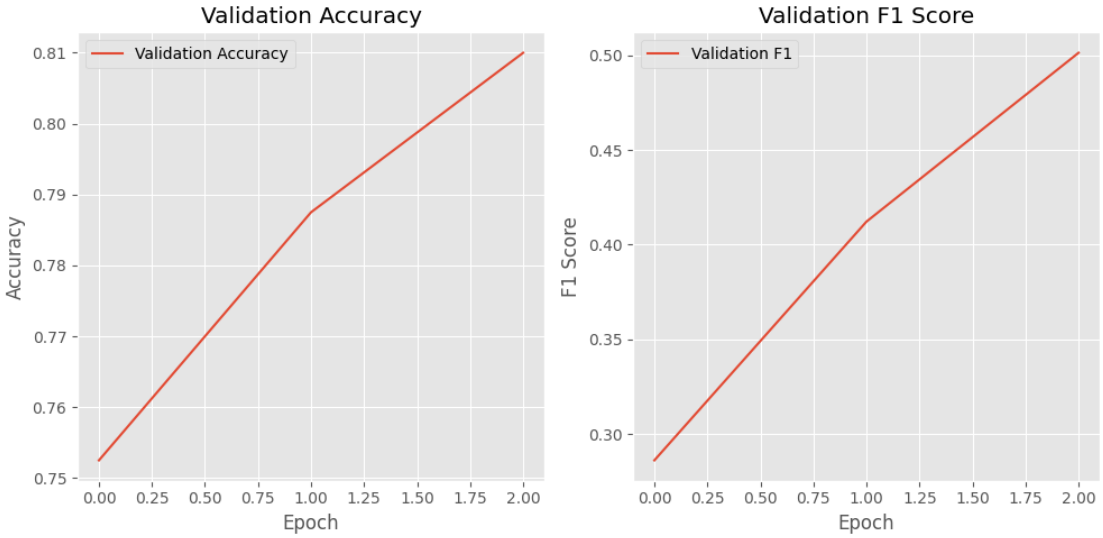

Figure 5.11: Accuracy & F1-score Validation

CHAPTER 6

Conclusion

6.1 Research Question 1

How efficiently can the proposed fine-tuned Sentiment Analysis model handle complex language and context in e-commerce data, such as sarcasm, to improve the accuracy of sentiment identification?

Our thesis shows that our fine-tuned RoBERTa model achieves excellent performance and is able to effectively handle sarcasm. We were able to design our model in such a way that it could maintain balanced performance across all categories, even though the distribution of sentiment classes in our dataset was uneven (as shown in our EDA in chapter 4.). This is reflected by the macro F1-score of 0.865, which indicates a strong precision and recall across all sentiment categories/classes.

Furthermore, our proposed model was able to achieve a 95.2% accuracy and an F1-score of 0.865, which outperformed both baseline models (VADER and RoBERTa). It also had particularly strong performance in neutral sentiment detection, an improvement from the baseline F1-score of 0.479 to 0.865, indicating an excellent understanding of nuanced expressions. In addition to these results, we also achieved near-perfect AUC scores across all 3 sentiment categories (Negative: 0.994, Neutral: 0.959, Positive: 0.993). Lastly, as far as sarcastic reviews are concerned, we significantly outperform the 2 base models, by achieving an Accuracy of 0.919, an F1-score of 0.903, and a Confusion Rate of 0.054.

6.2 Research Question 2

Can the model be extended to support multi-lingual and cross-lingual sentiment analysis, enabling businesses to analyze customer sentiments across different languages and cultural

contexts?

Addressing Research Question 2 is more challenging than the first one. While our extended model shows basic multilingual capability in 4 non-English languages, it falls short of the monolingual fine-tuned model's performance. This indicates that true cross-lingual sentiment understanding remains a challenge that requires further research and development.

The accuracy scores across all the tested languages that we were able to achieve were: Italian (84%), German (83.2%), Spanish (80.0%), and French (76.8%). But on the other hand, there is a performance gap compared to our fine-tuned model (for multilingual purposes), where the accuracy drops by 13-21%. F1-scores are even more noticeable, since they range from 0.464 to 0.530, which is lower than the monolingual model's 0.942. This performance variation across language shows some underlying challenges. First, by using a translation-based approach through Google Translate APIs, the model might introduce noise or lose sentiment indicators. Second, with the current architecture of our model we might not be able to capture cultural or linguistic differences in sentiment expression. Third, as the low F1-scores are clear evidence, the model struggles more with maintaining balanced precision and recall across languages. This comes as an outcome due to the fact that we use a translation-based approach through Google Translate API. The reason for this is because this approach is free and fits our small-scale real-world application. Gaining access to large datasets in different languages is challenging, as companies like Amazon or Alibaba don't share this data for free.

Our quantitative results in Chapter 5 show both the success of our approach in monolingual sentiment analysis and the remaining challenges in multilingual extension. This outcome provides valuable insights - as we see in Figures 5.11, 5.13, and 5.14 - for future work and research directions in cross-lingual sentiments, which highlight the complexity of building language-agnostic Sentiment Analysis models and systems.

6.3 Contributions

Through our model fine-tuning and the results we achieved, we can already contribute to the field of Sentiment Analysis in E-commerce. More specifically, we were able to:

- Develop a high-performance fine-tuned model specifically optimized for small-scale e-commerce reviews, achieving an accuracy of 95.2%.
- Develop a robust fine-tuning methodology that improves neutral sentiment detection, a traditionally challenging aspect of models like the VADER model.
- Provide empirical evidence of performance trade-offs in multilingual sentiment analysis.

- Establish baseline performance metrics for multilingual sentiment analysis in e-commerce contexts.

6.4 Future Work

Our research and model fine-tuning, especially the Multilingual model, reveals multiple promising directions for future research and contribution. They demonstrate strong performance, but are still prone to certain limitations and we believe there is still room for further improvement. The field of Sentiment Analysis in E-Commerce could heavily benefit from answering or addressing these challenges through extensive research, which could potentially lead to more robust and better performing SA models. More notably, here are some of our current limitations and future directions:

Multilingual Performance Gap

- Current limitation: Significant drop in performance for non-English languages
- Future direction: Investigate language-specific fine-tuning and cross-lingual knowledge transfer techniques

Cultural Nuance

- Current limitation: Limited understanding of culture-specific expressions
- Future direction: Develop culture-aware sentiment analysis incorporating regional linguistic patterns

Real-time Processing

- Current limitation: Computational overhead in multilingual processing
- Future direction: Research efficient model compression and optimization techniques

Aspect-based Analysis

- Current limitation: Global sentiment detection without aspect-specific analysis, which refers to the ability to identify and analyze sentiments directed at specific features or characteristics of a product
- Future direction: Extend the model to identify sentiment towards specific product aspects

List of Figures

List of Tables

Bibliography

[AAAA19] Shakeel Ahmad, Muhammad Zubair Asghar, Fahad M Alotaibi, and Irfanullah Awan. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Human-centric Computing and Information Sciences*, 9:1–23, 2019.

[AAK+22] Alim Al Ayub Ahmed, Sugandha Agarwal, IMade Gede Ariestova Kurniawan, Samuel PD Anantadjaya, and Chitra Krishnan. Business boosting through sentiment analysis using artificial intelligence approach. *International Journal of System Assurance Engineering and Management*, 13(Suppl 1):699–709, 2022.

[Ali12] Ahmed H Aliwy. Tokenization as preprocessing for arabic tagging system. *International Journal of Information and Education Technology*, 2(4):348, 2012.

[AM20] Shivaji Alaparthi and Manit Mishra. Bidirectional encoder representations from transformers (bert): A sentiment analysis odyssey. *arXiv preprint arXiv:2007.01127*, 2020.

[ASKP20] J Anvar Shathik and K Krishna Prasad. A literature review on application of sentiment analysis using machine learning techniques. *Int J Appl Eng Manag Lett (IJAEML)*, 4(2):41–67, 2020.

[ASQAA+18] Mohammad Al-Smadi, Omar Qawasmeh, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels' reviews. *Journal of computational science*, 27:386–393, 2018.

[AVG23] T Aravindan, C Vigneshwar, and Suganeshwari Ga. Sentiment classification for amazon fine foods reviews using pyspark. In *Recent Developments in Electronics and Communication Systems: Proceedings of the First International Conference on Recent Developments in Electronics and Communication Systems (RDECS-2022)*, volume 32, page 431. IOS Press, 2023.

[BAB21] TK Balaji, Chandra Sekhara Rao Annavarapu, and Annushree Bablani. Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, 40:100395, 2021.

[BAP+20] Yahia Baashar, Hitham Alhussian, Ahmed Patel, Gamal Alkawsi, Ahmed Ibrahim Alzahrani, Osama Alfarraj, and Gasim Hayder. Customer relationship management systems (crms) in the healthcare environment: A systematic literature review. *Computer Standards & Interfaces*, 71:103442, 2020.

[BDRS20] Rajesh Bose, Raktim Kumar Dey, Sandip Roy, and Debabrata Sarddar. Sentiment analysis on online product reviews. In *Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018*, pages 559–569. Springer, 2020.

[BJE15] Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599*, 2015.

[BKBH21] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134, 2021.

[BMS11] Cristina Bicchieri, Ryan Muldoon, and Alessandro Sontuoso. Social norms. 2011.

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[CD07] Padraig Cunningham and Sarah Delany. k-nearest neighbour classifiers. *Mult Classif Syst*, 54, 04 2007.

[CD21] Keith Cortis and Brian Davis. Over a decade of social opinion mining: a systematic review. *Artificial intelligence review*, 54(7):4873–4965, 2021.

[CDBF17] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. Affective computing and sentiment analysis. *A practical guide to sentiment analysis*, pages 1–10, 2017.

[CHPR+19] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*, 2019.

[CJJ+18] Eric M Clark, Ted James, Chris A Jones, Amulya Alapati, Promise Ukandu, Christopher M Danforth, and Peter Sheridan Dodds. A sentiment analysis of breast cancer treatment experiences and healthcare perceptions across twitter. *arXiv preprint arXiv:1805.09959*, 2018.

[CLC20] Liang-Chu Chen, Chia-Meng Lee, and Mu-Yen Chen. Exploration of social media for sentiment analysis using deep learning. *Soft Computing*, 24(11):8187–8197, 2020.

[CP21] Jonnathan Carvalho and Alexandre Plastino. On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. *Artificial Intelligence Review*, 54:1887–1936, 2021.

[CPE22] Rosario Catelli, Serena Pelosi, and Massimo Esposito. Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian. *Electronics*, 11(3):374, 2022.

[CXX+21] Zhaowei Chen, Yun Xue, Luwei Xiao, Jinpeng Chen, and Haolan Zhang. Aspect-based sentiment analysis using graph convolutional networks and co-attention mechanism. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part VI 28*, pages 441–448. Springer, 2021.

[DK19] Zulfadzli Drus and Haliyana Khalid. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161:707–714, 2019.

[DR23] Kerstin Denecke and Daniel Reichenpfader. Sentiment analysis of clinical narratives: A scoping review. *Journal of Biomedical Informatics*, 140:104336, 2023.

[DWT+20] Sanjay Dey, Sarhan Wasif, Dhiman Sikder Tonmoy, Subrina Sultana, Jayjeet Sarkar, and Monisha Dey. A comparative study of support vector machine and naive bayes classifier for sentiment analysis on amazon product reviews. In *2020 International Conference on Contemporary Computing and Applications (IC3A)*, pages 217–220. IEEE, 2020.

[EABH23] Anas El-Ansari and Abderrahim Beni-Hssane. Sentiment analysis for personalized chatbots in e-commerce applications. *Wireless Personal Communications*, 129(3):1623–1644, 2023.

[ENSN20] Christopher Ifeanyi Eke, Azah Anir Norman, Liyana Shuib, and Henry Friday Nweke. Sarcasm identification in textual data: systematic review, research challenges and open directions. *Artificial Intelligence Review*, 53:4215–4258, 2020.

[EXT+21] Ashkan Ebadi, Pengcheng Xi, Stéphane Tremblay, Bruce Spencer, Raman Pall, and Alexander Wong. Understanding the temporal evolution of covid-19 research through machine learning and natural language processing. *Scientometrics*, 126:725–739, 2021.

[FE19] Alessio Ferrari and Andrea Esuli. An nlp approach for cross-domain ambiguity detection in requirements engineering. *Automated Software Engineering*, 26(3):559–598, 2019.

[Fel99] Susan Feldman. Nlp meets the jabberwocky: Natural language processing in information retrieval. *ONLINE-WESTON THEN WILTON-*, 23:62–73, 1999.

[FGG97] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29:131–163, 1997.

[FHM08] Ingo Feinerer, Kurt Hornik, and David Meyer. Text mining infrastructure in r. *Journal of statistical software*, 25:1–54, 2008.

[FR22] Siavash Farzadnia and Iman Raeesi Vanani. Identification of opinion trends using sentiment analysis of airlines passengers' reviews. *Journal of Air Transport Management*, 103:102232, 2022.

[GE06] Isabelle Guyon and André Elisseeff. An introduction to feature extraction. In *Feature extraction: foundations and applications*, pages 1–25. Springer, 2006.

[GR19] Vaishali Ganganwar and R Rajalakshmi. Implicit aspect extraction for sentiment analysis: A survey of recent approaches. *Procedia Computer Science*, 165:485–491, 2019.

[HAM23] Huang Huang, Adeleh Asemi, and Mumtaz Mustafa. Sentiment analysis in e-commerce platforms: A review of current techniques and future directions. *IEEE Access*, PP:1–1, 01 2023.

[HCF12] Samuel T Hunter, Liliya Cushenbery, and Tamara Friedrich. Hiring an innovative workforce: A necessary yet uniquely challenging endeavor. *Human resource management review*, 22(4):303–322, 2012.

[HDO+98] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[HF17] Viktor Hangya and Richárd Farkas. A comparative empirical study on social media sentiment analysis over various genres and languages. *Artificial Intelligence Review*, 47:485–505, 2017.

[HLYJ19] Te Han, Chao Liu, Wenguang Yang, and Dongxiang Jiang. A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. *Knowledge-based systems*, 165:474–487, 2019.

[HLZZ24] Yufeng Han, Yang Liu, Guofu Zhou, and Yingzi Zhu. Technical analysis in the stock market: A review. *Handbook of Investment Analysis, Portfolio Management, and Financial Derivatives: In 4 Volumes*, pages 1893–1928, 2024.

[Hoc97] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[Hus18] Doaa Mohey El-Din Mohamed Hussein. A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 30(4):330–338, 2018.

[JSP21] Praphula Kumar Jain, Vijayalakshmi Saravanan, and Rajendra Pamula. A hybrid cnn-lstm: A deep learning approach for consumer sentiment analysis using qualitative user-generated contents. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–15, 2021.

[JSV21] Bineet Kumar Jha, GG Sivasankari, and KR Venugopal. Sentiment analysis for e-commerce products using natural language processing. *Annals of the Romanian Society for Cell Biology*, pages 166–175, 2021.

[JTM+24] Jamin Rahman Jim, Md Apon Riaz Talukder, Partha Malakar, Md Mohsin Kabir, Kamruddin Nur, and M.F. Mridha. Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, 6:100059, 2024.

[JYPS21] Praphula Kumar Jain, Ephrem Admasu Yekun, Rajendra Pamula, and Gautam Srivastava. Consumer recommendation prediction in online reviews using cuckoo optimized machine learning models. *Computers and Electrical Engineering*, 95:107397, 2021.

[KDS20] Olivier Kraaijeveld and Johannes De Smedt. The predictive power of public twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, 65:101188, 2020.

[KJ23] Diptesh Kanojiaa and Aditya Joshib. Applications and challenges of sa in real-life scenarios. *arXiv preprint arXiv:2301.09912*, 2023.

[KKKS22] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 82, 07 2022.

[KKKS23] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023.

[KKO+23] Arodh Lal Karn, Rakshha Kumari Karna, Bhavana Raj Ondamudi, Girish Bagale, Denis A Pustokhin, Irina V Pustokhina, and Sudhakar Sengan. Retracted article: Customer centric hybrid recommendation system for e-commerce applications by integrating hybrid sentiment analysis. *Electronic commerce research*, 23(1):279–314, 2023.

[KS23] Gagandeep Kaur and Amit Sharma. A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *Journal of big data*, 10(1):5, 2023.

[KYR19] Sudhanshu Kumar, Mahendra Yadava, and Partha Pratim Roy. Fusion of eeg response and sentiment analysis of products review to predict customer satisfaction. *information fusion*, 52:41–52, 2019.

[Las84] Roger Lass. *Phonology: An introduction to basic concepts*. Cambridge University Press, 1984.

[LCT21] Alexander Ligthart, Cagatay Catal, and Bedir Tekinerdogan. Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, pages 1–57, 2021.

[LH09] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384, 2009.

[Liu22] Bing Liu. *Sentiment analysis and opinion mining*. Springer Nature, 2022.

[LSG+22] Bin Liang, Hang Su, Lin Gui, Erik Cambria, and Ruifeng Xu. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, 235:107643, 2022.

[LZ12] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer, 2012.

[MAK20] Mohammad Erfan Mowlaei, Mohammad Saniee Abadeh, and Hamidreza Keshavarz. Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Systems with Applications*, 148:113234, 2020.

[MAK+23] Junaid Maqbool, Preeti Aggarwal, Ravreet Kaur, Ajay Mittal, and Ishfaq Ali Ganaie. Stock prediction by integrating sentiment scores of financial news and mlp-regressor: A machine learning approach. *Procedia Computer Science*, 218:1067–1078, 2023. International Conference on Machine Learning and Data Engineering.

[MAY+24] Erkut Memiş, Hilal Akarkamçı, Mustafa Yeniad, Javad Rahebi, and Jose Manuel Lopez-Guede. Comparative study for sentiment analysis of financial tweets with deep learning methods. *Applied Sciences*, 14(2):588, 2024.

[MCK21] Fuad Mehraliyev, Irene Cheng Chu Chan, and Andrei Kirilenko. Sentiment analysis in hospitality and tourism: a thematic and methodological review. *International Journal of Contemporary Hospitality Management*, ahead-of-print, 10 2021.

[Mcl23] S Mcleod. Social identity theory in psychology (tajfel & turner, 1979). simply psychology, 2023.

[MEKCP14] Daniel McDuff, Rana El Kaliouby, Jeffrey F Cohn, and Rosalind W Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing*, 6(3):223–235, 2014.

[MHK14] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[MJ+01] Larry R Medsker, Lakhmi Jain, et al. Recurrent neural networks. *Design and Applications*, 5(64-67):2, 2001.

[MMC15] Tim K Mackey, Angela Miner, and Raphael E Cuomo. Exploring the e-cigarette e-commerce marketplace: Identifying internet e-cigarette marketing characteristics and regulatory gaps. *Drug and alcohol dependence*, 156:97–103, 2015.

[MWW+18] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[NKU21] KE Naresh Kumar and V Uma. Intelligent sentinet-based lexicon for context-aware sentiment analysis: optimized neural network for sentiment classification on social media. *The Journal of Supercomputing*, 77(11):12801–12825, 2021.

[OADAH21] Ruba Obiedat, Duha Al-Darras, Esra Alzaghoul, and Osama Harfoushi. Arabic aspect-based sentiment analysis: A systematic literature review. *IEEE Access*, 9:152628–152645, 2021.

[O'S15] K O'Shea. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[OV11] Mihály Ormos and Miklós Vázsonyi. Impacts of public news on stock market prices: Evidence from sp5001. *Interdisciplinary Journal of Research in Business*, 1:1–17, 03 2011.

[Par03] Daniel J Paré. Does this site deliver? b2b e-commerce services for developing countries. *The Information Society*, 19(2):123–134, 2003.

[Pat15] S Patro. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.

[PCZ+22] Sancheng Peng, Lihong Cao, Yongmei Zhou, Zhouhao Ouyang, Aimin Yang, Xinguang Li, Weijia Jia, and Shui Yu. A survey on deep learning for textual emotion analysis in social networks. *Digital Communications and Networks*, 8(5):745–762, 2022.

[PHC+18] Soujanya Poria, Amir Hussain, Erik Cambria, Soujanya Poria, Amir Hussain, and Erik Cambria. Combining textual clues with audio-visual information for multimodal sentiment analysis. *Multimodal sentiment analysis*, pages 153–178, 2018.

[PMS17] Rajesh Piryani, Devaraj Madhavi, and Vivek Kumar Singh. Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management*, 53(1):122–150, 2017.

[PPC20] Han Woo Park, Sejung Park, and Miyoung Chong. Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea. *Journal of medical internet research*, 22(5):e18897, 2020.

[R+01] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. Citeseer, 2001.

[RAA+22] G Revathy, Saleh A Alghamdi, Sultan M Alahmari, Saud R Yonbawi, Anil Kumar, and Mohd Anul Haq. Sentiment analysis using machine learning: Progress in the machine intelligence for data science. *Sustainable Energy Technologies and Assessments*, 53:102557, 2022.

[RD+00] Erhard Rahm, Hong Hai Do, et al. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[RHZ20] Lavinia Rognone, Stuart Hyde, and S Sarah Zhang. News sentiment in the cryptocurrency market: An empirical comparison with forex. *International Review of Financial Analysis*, 69:101462, 2020.

[RKK20] Nikolas Ruffer, Johannes Knitza, and Martin Krusche. # covid4rheum: an analytical twitter study in the time of the covid-19 pandemic. *Rheumatology International*, 40(12):2031–2037, 2020.

[RR21] Albérico Rosário and Ricardo Raimundo. Consumer marketing strategy and e-commerce in the last decade: A literature review. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(7):3003–3024, 2021.

[SAR23] Yadav Meenakshi Muthukrishnan Seethalakshmi, Suruliandi Andavar, and Raja Soosaimarian Peter Raj. A survey on feature extraction techniques, classification methods and applications of sentiment analysis. *Brazilian Archives of Biology and Technology*, 66:e23220654, 2023.

[SF15] Kim Schouten and Flavius Frasincar. Survey on aspect-level sentiment analysis. *IEEE transactions on knowledge and data engineering*, 28(3):813–830, 2015.

[SGJ+17] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

[Sha19] Rutvik Sharedalal. Amazon fine food reviews-design and implementation of an automated classification system. 2019.

[SL20] Raksmey Sann and Pei-Chun Lai. Understanding homophily of service failure within the hotel guest cycle: Applying nlp-aspect-based sentiment analysis to the hospitality industry. *International Journal of Hospitality Management*, 91:102678, 2020.

[SLZ+21] LDCS Subhashini, Yuefeng Li, Jinglan Zhang, Ajantha S Atukorale, and Yutong Wu. Mining and classifying customer reviews: a survey. *Artificial Intelligence Review*, pages 1–47, 2021.

[SPMM20] Santwana Sagnika, Anshuman Pattanaik, Bhabani Shankar Prasad Mishra, and Saroj K Meher. A review on multi-lingual sentiment analysis by machine learning methods. *Journal of Engineering Science and Technology Review*, 13(2):154, 2020.

[TBT+11] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[VGEVA19] Franco Valencia, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy*, 21(6):589, 2019.

[Wal13] Douglas Walton. *Fallacies arising from ambiguity*, volume 1. Springer Science & Business Media, 2013.

[WAV22] Mayur Wankhade, Chandra Sekhara Rao Annavarapu, and Mukul Kirti Verma. Cbvosd: context based vectors over sentiment domain ensemble model for review classification. *The Journal of Supercomputing*, pages 1–37, 2022.

[WWH09] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433, 2009.

[XCW18] Frank Z Xing, Erik Cambria, and Roy E Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73, 2018.

[YC14] Bishan Yang and Claire Cardie. Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 325–335, 2014.

[YCL+19] Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60:617–663, 2019.

[YQG+19] Mengbin Ye, Yuzhen Qin, Alain Govaert, Brian DO Anderson, and Ming Cao. An influence network model to study discrepancies in expressed and private opinions. *Automatica*, 107:371–381, 2019.

[YSS22] Sruthi Yarkareddy, T Sasikala, and S Santhanalakshmi. Sentiment analysis of amazon fine food reviews. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1242–1247. IEEE, 2022.

[YSY22] Jenny Yow Bee Yin, Nor Hasliza Md Saad, and Zulnaidi Yaacob. Exploring sentiment analysis on e-commerce business: Lazada and shopee. *Tem journal*, 11(4):1508–1519, 2022.

[YV20] Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385, 2020.

[ZAZC18] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, 9(2):1–29, 2018.

[ZLD+22] Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038, 2022.

[ZO18] Kudakwashe Zvarevashe and Oludayo O Olugbara. A framework for sentiment analysis with opinion mining of hotel reviews. In *2018 Conference on information communications technology and society (ICTAS)*, pages 1–4. IEEE, 2018.

[ZWL18] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.

[ZXW19] Yabing Zhao, Xun Xu, and Mingshu Wang. Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management*, 76:111–121, 2019.

[ZZCC20] Enguang Zuo, Hui Zhao, Bo Chen, and Qiuchang Chen. Context-specific heterogeneous graph convolutional network for implicit sentiment analysis. *IEEE Access*, 8:37967–37975, 2020.

[ZZZW20] Shaozhong Zhang, Dingkai Zhang, Haidong Zhong, and Guorong Wang. A multiclassification model of sentiment for e-commerce reviews. *IEEE Access*, 8:189513–189526, 2020.