

# **Evaluating Large Language Model Architectures for Sentiment Analysis Across Multiple Datasets**

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE AWARD OF THE DEGREE  
OF

MASTERS OF TECHNOLOGY  
IN  
**Information Technology**

Submitted by

**KSHITIJ PRAKASH SRIVASTAVA**  
**2K23/ITY/17**

Under the supervision of

**PROF. DINESH K. VISHWAKARMA**



**INFORMATION TECHNOLOGY**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi 110042

**MAY 2025**

**DEPARTMENT OF INFORMATION TECHNOLOGY**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi-110042

**CANDIDATE’S DECLARATION**

I, **Kshitij Prakash**, Roll Numbner – **2k23/ITY/17** students of M.Tech (**Information Technology**), hereby declare that the project Dissertation titled “**Evaluating Large Language Model Architectures for Sentiment Analysis Across Multiple Datasets**” which is submitted by me to the **Information Technology** Department, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Masters of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

**Kshitij Prakash Srivastava**

Date: 29.05.2025

**DEPARTMENT OF INFORMATION TECHNOLOGY**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi-110042

**CERTIFICATE**

I hereby certify that the Project Dissertation titled “**Evaluating Large Language Model Architectures for Sentiment Analysis Across Multiple Datasets**” which is submitted by **Kshitij Prakash Srivastava, 2k23/ITY/17, Information Technology**, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Masters of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

**Prof. Dinesh K. Vishwakarma**

Date: 29.05.2025

**SUPERVISOR**

**DEPARTMENT OF INFORMATION TECHNOLOGY**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi-110042

**ACKNOWLEDGEMENT**

We wish to express our sincerest gratitude to **Prof. Dinesh K. Vishwakarma** for his continuous guidance and mentorship that he provided us during the project. He showed us the path to achieve our targets by explaining all the tasks to be done and explained to us the importance of this project as well as its industrial relevance. He was always ready to help us and clear our doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

Place: Delhi

**Kshitij Prakash Srivastava**

Date:

# Abstract

What is now referred to as LLMs and their rapid growth have touched every sphere of Natural Language Processing. Therefore, sentiment analysis continues to be an application worth understanding as it provides insights into opinions, emotions, and attitudes being expressed in textual data. However, with the existence of LLMs, interpretation is still a challenge when it comes to subtle language use, complex linguistic phenomena such as sarcasm, and blatant issues of biases within models. The objective of this thesis is to analyse and assess the performance of different state-of-the-art LLMs in sentiment analysis on various datasets and under different sentiment paradigms. A full experimental setup for such analysis was developed, which uses a handful of major LLMs, including proprietary ones such as GPT-4o for benchmarking the cutting-edge performance and also opensource variants like Llama 3, Mistral Large/Mixtral-8x22B, Falcon LLM, and XLM-RoBERTa for their accessibility, transparency, and customizability. The experiments covered a variety of sentiment analysis tasks such as those for binary classification (IMDb Movie Reviews, SST-2), aspect-based sentiment analysis (MAMS-for-ABSA), multilingual sentiment analysis (Multilingual Amazon Reviews Corpus), and tests for harder cases such as sarcasm detection. To achieve this, a comprehensive experimental framework was developed, utilizing a selection of prominent LLMs, including proprietary models such as GPT-4o for benchmarking against cutting-edge performance, and open-source alternatives like Llama 3, Mistral Large/Mixtral-8x22B, Falcon LLM, and XLM-RoBERTa for their accessibility, transparency, and customizability. Experiments were conducted on a range of sentiment analysis tasks, with binary classification (IMDb Movie Reviews, SST-2), aspect-based sentiment analysis (MAMS-for-ABSA), multilingual sentiment analysis (Multilingual Amazon Reviews Corpus), and challenging scenarios such as sarcasm detection (Twitter Sentiment Analysis Datasets). The research explored the efficacy of zero-shot, few-shot, and fine-tuning approaches, emphasizing the critical role of prompt engineering in optimizing LLM performance. For sentiment analysis, older deep learning

architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs - including LSTMs) primarily focused on capturing local patterns (CNNs) or sequential dependencies (RNNs) within text.

# Contents

<b>Candidate’s Declaration</b>	<b>i</b>
<b>Certificate</b>	<b>ii</b>
<b>Acknowledgement</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>Content</b>	<b>vii</b>
<b>List of Tables</b>	<b>viii</b>
<b>List of Figures</b>	<b>ix</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Motivation . . . . .	2
1.2 Problem Statement . . . . .	5
1.3 Overview . . . . .	5
1.4 Research Gaps . . . . .	6
1.5 Filling Research Gaps . . . . .	6
<b>2 LITERATURE REVIEW</b>	<b>9</b>
2.1 Natural Language Processing . . . . .	9
2.2 Traditional Approach . . . . .	9
2.3 Sentiment Analysis . . . . .	11
2.4 Transformers . . . . .	12
2.4.1 Tokens . . . . .	14
2.4.2 Embeddings . . . . .	15
<b>3 METHODOLOGY</b>	<b>17</b>
3.1 Datasets used . . . . .	19
3.1.1 Yelp-Polarity . . . . .	19
3.1.2 IMDB Movie Reviews . . . . .	20
3.1.3 Stanford Sentiment Treebank v2 (SST-2) . . . . .	21
3.1.4 Multilingual Amazon Reviews Corpus (MARC) . . . . .	22
3.2 Large Language Models Used . . . . .	24
3.2.1 BERT . . . . .	24
3.2.2 GPT-4o . . . . .	25
3.2.3 Text-to-Text-Transfer-Transformer (T5) . . . . .	26
3.2.4 XLM-RoBERTa . . . . .	27

<b>4</b>	<b>RESULTS and DISCUSSION</b>	<b>29</b>
4.1	Performance on Multi-Lingual Dataset . . . . .	29
4.1.1	Comparative Analysis on MARC . . . . .	30
4.2	Performance on Binary Datasets . . . . .	33
<b>5</b>	<b>CONCLUSION AND FUTURE SCOPE</b>	<b>36</b>
5.1	Conclusion . . . . .	36
5.2	Future Scope . . . . .	36



## List of Tables

3.1	Yelp Polarity Dataset Structure . . . . .	20
3.2	IMDB Movie Reviews Dataset Structure . . . . .	21
3.3	Stanford Sentiment Treebank v2 (SST-2) Dataset Structure . . . . .	22
3.4	Multilingual Amazon Reviews Corpus (MARC) Dataset Structure . . . . .	23
3.5	Comparison of BERT, GPT-4o, Llama-3, and XLM-RoBERTa . . . . .	28
4.1	Model Performance Metrics . . . . .	29
4.2	Performance Metrics of LLMs on multiple Datasets . . . . .	33

## List of Figures

1.1	Subcategories of NLP . . . . .	2
2.1	Recurrent Neural Network Architecture . . . . .	10
2.2	Long Short Term Memory Model . . . . .	10
2.3	Sentiment Classification . . . . .	11
2.4	Transformers Development Timeline . . . . .	12
2.5	Key-Query Space . . . . .	13
2.6	Tokenization . . . . .	14
2.7	Embedding Process . . . . .	15
3.1	Sentiment Classification Pipeline . . . . .	18
3.2	Bidirectional Encoder Representations from Transformers . . . . .	25
4.1	Accuracy Comparison . . . . .	30
4.2	Precision Comparison . . . . .	31
4.3	Recall Comparison . . . . .	31
4.4	F1-Score Comparison . . . . .	32
4.5	F1-Score Comparison . . . . .	34
4.6	Metrics Comparisons . . . . .	35

# Chapter 1

## INTRODUCTION

The natural language processing and user-interfacing fields are ever-changing, and with the emergence of large language models (LLMs), text analysis has now reached a deeper and more accurate level. From the pre-2010 days scaling through to the evolution of large language models, the task was primarily handled by rules-based systems and simple ML models like Naive Bayes or Support Vector Machines (SVMs). These did not do well with context and fine nuances around sentiments. The introduction of word embeddings (Word2Vec in 2013, followed by GloVe in 2014) helped with semantic relations but not with contextual relations. The biggest advancement began with transformers-based models, starting from 2018 with Google BERT, which utilized bidirectional context for sentiment capture [6]. BERT’s huge success led to furious developments: OpenAI’s GPT-2 (2019) and GPT-3 (2020) brought the few-shot learning revolution to sentiment analysis, where the systems could work with a few labels [3]. By 2021, RoBERTa and XLNet, furthering BERT, outperformed it on sentiment tasks [12][33]. Then came the rise of ChatGPT in 2022 and GPT-4 in 2023, another big jump wherein the models started inferring subtle emotions, sarcasm, and various cultural nuances with great precision [47]. The new models like Meta’s LLaMA (2023) and fine-tuned open-source models (e.g., Mistral, Falcon) have democratized quality sentiment analysis [25][24].

Together with its rapid evolution, LLMs have brought forth new problems and opportunities facing sentiment analysis. Early models had a hard time with domain adaptation, with newer architectures such as Google’s T5 and OpenAI’s GPT-4 displaying superb zero-shot and few-shot abilities, thereby generalizing across industries—from healthcare to finance—without need for additional training [17][47]. The dawn of instruction-tuned models, such as Alpaca and Vicuna, has introduced a greater degree of intuitive fine-tuning, where sentiment classification is further refined through human feedback. In this setting, multimodal LLMs, such as GPT-4V and Gemini, meanwhile, bring in the visual modality in addition to textual cues, boosting sentiment analysis in social media posts involving an interplay of images and text [40]. Another key aspect toward achieving unbiased AI is consideration of such sort of methods that render bias mitigationability onto sentiment analysis,

thereby averting distortions in interpretations, (bias elimination) through models such as Claude by Anthropic. Regarding further alternatives, there is Zephyr and StableLM, which provide for a custom-made sentiment analysis method, allowing businesses to customize it towards their niche datasets. So, with the ever-growing power of LLMs, real-time applications, such as live customer feedback analysis and stock market sentiment tracking, are getting more and more accurate [45][49]. This, therefore, states that the future of sentiment analysis lies in adaptive, context-aware AI systems.

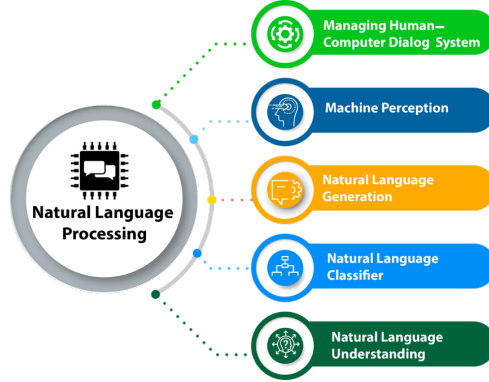


Figure 1.1: Subcategories of NLP

Today, LLMs leverage multimodal inputs (text, audio, and visuals) for richer sentiment detection, while techniques like prompt engineering and chain-of-thought reasoning further refine outputs [40]. Despite challenges like bias and computational costs, LLMs have revolutionized NLP, moving from rigid keyword matching to dynamic, human-like interpretation [48][49]. Future directions may focus on real-time analysis, ethical AI, and lightweight models for broader adoption [50][42].

## 1.1 Motivation

Rapid evolution of large language models (LLM) has fundamentally altered natural language processing (NLP) with respect to generating raw texts, extracting sentiments, or summarizing information, among others. With an increasing trend of LLMs proliferating the scene, from free offerings such as BERT and GPT-2 to the commercial giants like GPT-4 and Claude [6][3][47], it becomes pertinent to analyze the performance of LLMs under different evaluation standards. Performing an LLM comparison using well-recognized datasets of Hugging Face (such as SST-2, Amazon Reviews, and IMDB) holds several important aspects. Among them are understanding the model’s strengths and weaknesses, further testing their generalization and robustness, studying the trade-offs made between efficiency and accuracy, keeping track of progress in NLP, helping boost transparency and repro-

ducibility, aiding strategic decisions within industrial and academic sectors, and pointing toward promising remit areas for the next deep dive [9][49]. A component of NLP, sentiment analysis seeks to detect and classify subjective information—such as opinions, feelings, and attitudes—into text with an assigned polarity, commonly positive, negative, or neutral [11][31]. From market research to competitor analysis, from interpretation of customer views to tracking brand perception, indeed, sentiment analysis has far-reaching applications, all the way to assessing public views in politics.

A key component of NLP, sentiment analysis focuses on identifying and classifying subjective data—like opinions, feelings, and attitudes—within text, usually assigning a positive, negative, or neutral polarity [11][31]. The impact of sentiment analysis is far-reaching, from market research and competitor analysis to interpreting customer opinions, tracking brand perception, and even examining public sentiment in politics. Different LLMs are optimized for particular functions; some perform exceptionally well in sentiment analysis, while others are better suited for text generation or question answering [34]. Evaluating these models on standardized datasets like SST-2 (for binary sentiment) or IMDB (for movie reviews) helps clarify which architectural designs (e.g., encoder-based models like BERT versus decoder-based models like GPT) are most effective for specific applications [9][6][3]. For instance, does RoBERTa provide more accurate sentiment analysis than DistilBERT for complex Amazon reviews? [12][50] Can a smaller model like Alpaca truly rival the performance of larger models after targeted fine-tuning? A methodical comparison helps practitioners select the ideal model.

In turn, LLMs outperform other models on such datasets and, thus, better reflect their ability to tackle language, contextual, and domain-specific sentiment variations [17][27]. For instance, while GPT-4 may shine in open-ended text generation tasks, in fine-grained sentiment classification—such as sarcasm detection or the analysis of mixed sentiments—it may lose out to domain-specific models like FinBERT, fine-tuned for financial texts [45]. Similarly, multilingual models such as XLM-RoBERTa can be assessed for cross-lingual sentiment analysis to check if their capabilities decay when it comes to low-resource languages [5][8][28]. Another factor worth considering is computational efficiency: models such as DistilBERT or MobileBERT may afford a slight trade-off in accuracy in return for faster inference, which is paramount for real-time tasks like social media monitoring [50].

In addition, performance perception differs widely according to the choice of evaluation metrics, whether accuracy, F1-score, or AUC-ROC. Compared to single-task models, a model like T5, with its multi-task training objective, is hypothesized to generalize better across tasks in sentiment analysis but depends on how complex the dataset is [17][30]. Comparative studies also shed light on the effect of fine-tuning, data augmentation, and prompt engineering on model performance [22][39]. For instance, few-shot learning with GPT-3.5 would compete well with little to

no need for heavy retraining [3], whereas BERT-based models often require much training [6].

## 1.2 Problem Statement

Testing models across multiple datasets (e.g., SST-2 for short phrases, IMDB for lengthy reviews) reveals whether a model’s performance is consistent or deteriorates with varying text lengths, domains, or linguistic complexity [6][9][17]. For example, does GPT-3.5 maintain high accuracy on both formal news data and informal social media text? Comparative analysis uncovers biases, overfitting tendencies, and robustness gaps that may not be apparent when evaluating a single model in isolation [20][39].

The NLP field evolves rapidly, with new architectures and training techniques emerging frequently. By benchmarking models on established datasets, we can track progress over time—does Llama 2 outperform its predecessor? How does Falcon compare to open-source alternatives? [13][43] A comparative analysis provides empirical evidence of advancements, helping researchers identify which innovations (e.g., better attention mechanisms, improved tokenization) contribute most to performance gains [25][37]. Many LLMs are released with claimed benchmarks, but independent verification is essential. Reproducing results on public datasets like those in the Hugging Face ecosystem ensures transparency and builds trust in model capabilities [6][30]. Additionally, inconsistencies in evaluation methodologies (e.g., different preprocessing steps, hyperparameters) can skew comparisons. A standardized analysis mitigates this by applying uniform evaluation criteria across all models [39][40]. Businesses and researchers must make informed choices when selecting models for deployment or further development.

By analyzing where models struggle (e.g., handling sarcasm in reviews, long-range dependencies in text), we highlight areas needing innovation [45]. If multiple models underperform on Amazon Reviews due to domain-specific jargon, this signals a need for better domain adaptation techniques [47]. Such insights guide future research directions.

## 1.3 Overview

A systematic evaluation of LLMs across Hugging Face datasets is not just an academic exercise—it’s a necessity for advancing NLP in a structured, efficient, and transparent manner [6][9][47]. By comparing models on accuracy, speed, robustness, and scalability, we empower developers, businesses, and researchers to leverage the right tools for their needs while driving the field toward more capable and accessible language technologies [27][39]. This comparative analysis lays the groundwork for better model selection, improved benchmarking standards, and targeted innovations in AI [48].

A defining characteristic of LLMs is their exhibition of ”emergent abilities,” which were not explicitly programmed but arise from their scale and training [3][26].

These abilities include in-context learning, where LLMs can learn a new task from a small set of examples provided within the prompt at inference time, and instruction following, allowing them to perform new types of tasks based solely on instructions without explicit examples [27][33]. Furthermore, LLMs can be augmented with external knowledge and tools, enhancing their capacity to interact with users and environments effectively [40][51].

## 1.4 Research Gaps

Despite the rapid advancements in Large Language Models, several research gaps persist in their evaluation and comparative analysis [37][38][42]. Many studies focus on benchmark performance but overlook real-world generalization, particularly in handling domain shifts, noisy data, and multilingual contexts [18][28][41]. Few evaluations rigorously assess computational efficiency, ethical biases, or explainability across models [36][48][49]. Additionally, most benchmarks use static datasets, failing to capture dynamic, evolving language trends [39][43]. There is also limited research on the trade-offs between fine-tuning efficiency and zero-shot capabilities [10][27]. Finally, standardized evaluation frameworks for comparing open-source and proprietary models are lacking, creating inconsistencies in performance claims [14][24]. Addressing these gaps would enable more robust, fair, and practical LLM assessments [38][48].

## 1.5 Filling Research Gaps

Comparative analysis of large language models (LLMs) across different text datasets such as product reviews, social media comments, and forum discussions—is crucial for understanding their strengths, limitations, and real-world applicability [4][9][18][26]. Every covert linguistic challenge posed by each dataset, such as attitude changes, jargon, or cultural context, might affect the model’s performance [5][7][30]. For example, sentiment analysis performed on product reviews may achieve a high level of accuracy, as the language is well structured and formal [15][20]. In contrast, social media comments provide an informal setting fractured by the use of slang, sarcasm, or emojis where a model may get tested in terms of understanding context [6][13]. Through systematic evaluation of LLMs—such as GPT-4, BERT, and LLaMA—over all these disciplines, researchers can identify biases, contradictions, or areas where the model fails to adapt that probably wouldn’t stem from conventional benchmarks [14][28][38]. One of the most important values added by the cross-comparison is exposing the architectural and methodological issues affecting performance [16][18][23]. Transformer family models fine-tuned on big sets of documents like RoBERTa tend to perform very well on generic tasks but might not work as well in focused domain scenarios where smaller targeted methods surpass them [10][25]. To pinpoint one, an LLM predominantly trained on



news probably gets lost in picking up on colloquial nuances in YouTube comments, whereas a model fine-tuned on user-generated content would excel at it [12][27]. Furthermore, across-language or dialectal comparisons also shed light on a more critical aspect of multilingual support needed for worldwide applications [17][19]. As biased datasets favor stereotypes during their social media discourse-phase dictation, a balanced training set can provide a reduction in discriminatory outcomes [36][48]. Disclosure of these model comparisons empowers stakeholders—businesses, policymakers, and developers—in the choice of “right” tool for a particular application [22][24]. For example, a company trying to deploy chatbots for customer service would be concerned with models with proven track record on conversational datasets more than those on formal documents [34]. As LLMs enter the next stage of maturity, continuing comparative research guarantees a smooth transition of advancements into working, fair, and scalable solutions; those solutions in turn will mold the future of NLP applications [38][42]. Even so, FinBERT still would be better over generic LLMs for financial text, thus, making a strong case for dedicated solutions [15].

Next, we have efficiency in computation working against them. DistilBERT allows for a 40% speedup over BERT by compromising a little on accuracy, which is beneficial for real-time processing [31]. Meanwhile, with the blossoming of zero-shot competences in such models as GPT-3.5, limitations imposed by dependence on labeled training data are less severe, although problems in implementation may arise concerning any language other than English, bringing in stark notice the lack in multilingual support [34]. Bias removal is another challenge: through HateCheck tests, even SOTA LLMs proved to be not without racial and gender bias, thus highlighting the need for vigorous audits for fairness [40]. When comparing large language models, for instance, GPT-4, BERT, and LLaMA, it is very important to test them on various styles of text—like product reviews, social media comments, and forums [4][9][18]. This is because every text type has its own set of complicated language features. Sometimes slang and emojis appear in social media text, which perplexes the models; meanwhile, product reviews are in clear and formal language, so the models fare better [6][15]. By inspecting these differences into finer detail, we get to give a boost to understanding where these models do well and where they lag [14][28]. For example, some models may be more trained on news articles and might just not get the casual talks on YouTube comments [12][27]. Also, certain models that are trained on really large and generic datasets would be fine for broader tasks, but in case of topics, specialized, smaller models might be the better choice [10][25]. And in the process of testing how these models perform with different languages or dialects, we end up realizing how well those models support multilingual users across the world [17][19]. This then helps in eliminating bias or stereotypes, which sometimes arise if the data on which the model learns is not balanced [36][48]. These details, in turn, help companies and developers to really find the right model for their use-case; for instance, if a company wants to develop a chatbot, they should opt for a model with good conversational skills instead of

one that performs well on formal writing [22][34]. So this very careful comparison of these language models ensures that the technology continues to improve and to remain fair and useful in the real world [38][42].

## Chapter 2

# LITERATURE REVIEW

## 2.1 Natural Language Processing

Natural language processing has witnessed drastic changes: it used to be rule-based; today, it is AI-driven modeling for human text comprehension. In the beginning, NLP employed handcrafted linguistic rules accompanied by statistical methods, which could hardly be scaled up or even accurate (Jurafsky and Martin, 2009). Then the big break came with the machine learning algorithms such as Hidden Markov Models (HMMs), Support Vector Machines (SVMs), etc., giving a new meaning to tasks like part-of-speech tagging and named entity recognition with a sense of precision. But these methods still could not cope well with ambiguity and context. The true revolution started with transformer architectures, or rather with their introduction by Vaswani et al. (2017), which did away with sequential processing channels and instead employed parallelizable attention mechanisms. This paved the way for pre-trained language models such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) that harnessed the power of huge amounts of data for their groundbreaking application to NLP tasks in general, ranging from sentiment analysis to machine translation. Such models have demonstrated zero-shot and few-shot method abilities on an unprecedented scale and thus have reduced dependence on task-specific training data. With expediency and ethics at the forefront, more recent models like LLaMA (Touvron et al., 2023) and Mistral consider addressing biases and optimizing for limited resources. Today, NLP, stitching threads that bind virtual assistants and real-time translation applications together, pushes hard on the interaction limits between humans and a machine.

## 2.2 Traditional Approach

Before the arrival of modern LLMs, sentiment analysis mainly used either RNNs or LSTMs to process texts like movie reviews, product feedback, and social media comments (Liu et al., 2016). Among the earliest neural architectures, RNNs featured sentiment classification in that they take sequential data, retaining a hidden

state that stores contextual information over time (Socher et al., 2013).

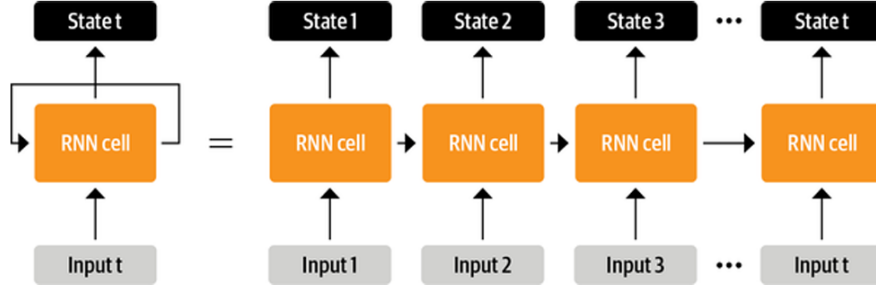


Figure 2.1: Recurrent Neural Network Architecture

Nevertheless, vanilla RNNs suffered from the vanishing-gradient problems, limiting retaining long-range dependencies inherent in a text (Bengio et al., 1994). An LSTM solves this problem by introducing gating mechanisms: input, forget, and output gates—that control the flow of information; hence, the LSTM could better model sentiment in long or complex reviews (Hochreiter Schmidhuber, 1997).

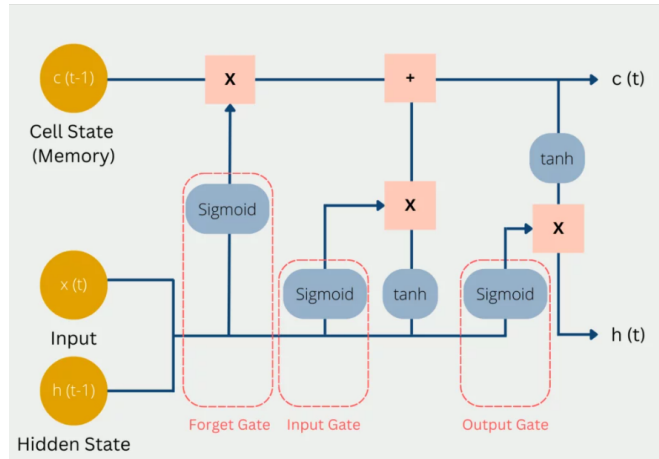


Figure 2.2: Long Short Term Memory Model

It was demonstrated that LSTMs performed better than more traditional machine learning methods such as SVMs and Naïve Bayes for sentiment analysis because these networks were able to detect more subtle contextual cues (Tang et al., 2015). Admittedly, LSTMs are computationally demanding and have difficulties with parallelizing; hence, they lack efficiency compared to later transformer-based models (Vaswani et al., 2017). Hybrid methods, in others, combine LSTMs with attention mechanisms, allowing for better sentiment classification by focusing on words that were important for expressing sentiment (Yang et al., 2016).

## 2.3 Sentiment Analysis

Sentiment Analysis, also called Opinion Mining, is a computational technique that identifies and categorizes emotions, a person's attitude, and opinions expressed in text data [1][3]. Through NLP and machine learning, the system can recognize whether a batch of written text is a positive remark, a reproachful statement, or a neutral comment [7][11]. Instances include companies analyzing social media posts or product reviews in order to grasp public perception in real-time [5][9]. Considerations are even given for context, with advanced models also trying to comprehend sarcasm, irony, or context-related tones; accuracy, however, is dependent on training data quality themselves [13][21]. One finds applications in finance (predicting market trends), healthcare (gauging patient feedback), and politics (polling public sentiments) [16][20][23].

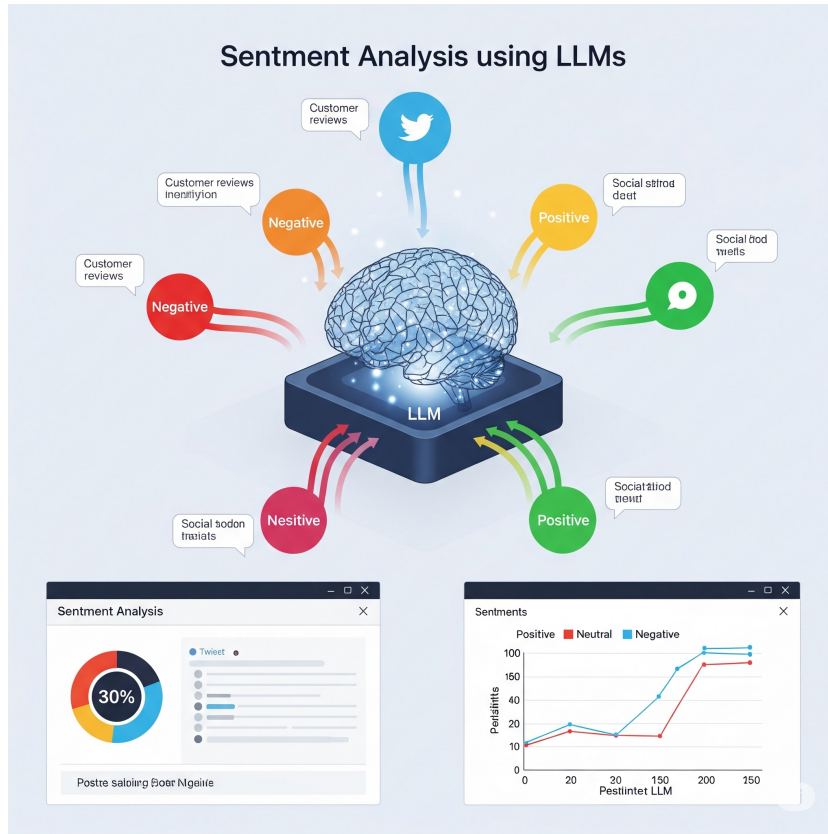


Figure 2.3: Sentiment Classification

It has its drawbacks such as using ambiguous language, opting for cultural nuances, and dealing with multilingual texts [13][14][15]. Nonetheless, it is worth mentioning that despite its set of shortcomings, analyzing context is infinitely valuable in making data-driven decisions because of its immense capability to process

huge datasets within a fraction of seconds [7][10]. Thanks to AI advancement, sentiment analysis gets deeper in understanding human feelings and behaviors [3][17]. In marketing, research, or customer service, this mechanism takes unstructured text data and gives it useful analytic output, thus providing a solution between raw data and significant meaning [5][8][12].

## 2.4 Transformers

The Transformer architecture, introduced by Vaswani et al. in 2017, has been an AI revolution in enabling parallelized attention mechanism architectures for the processing of sequential data, thereby making it incredibly suitable for sentiment analysis [2]. Whereas recurrent neural networks (RNNs) try to assign importance to words in a sentence by virtue of sequential order—with the self-attention mechanism of Transformers it is possible to assign such weights depending on other criteria on a much longer, if not the longest, potential scale; this becomes important for sentiment-specific long-range dependencies [6]. For sentiment analysis purposes, models like BERT (Bidirectional Encoder Representations from Transformers) fine-tune some pre-trained Transformer layers in order to determine whether the given text is classified as positive, negative, or neutral by analyzing the contextual relations between words [19]. The multi-head attention mechanism of this architecture is able to attend to various sentiment indicators simultaneously and thus perform better compared to classical sentiment classification techniques [20]. Transformer-based methods have been proven to achieve state-of-the-art performance on sentiment analysis benchmarks like IMDb and SST-2, even outperforming the older techniques by great margins [21]. Although very computationally expensive, they need to be optimized by distilling or pruning before practical implementation [22]. By utilizing attention mechanisms to excel in the detection of subtle sentiment such as sarcasm or mixed feelings, the Transformers thus become the cornerstone of a modern natural language processing application [7].

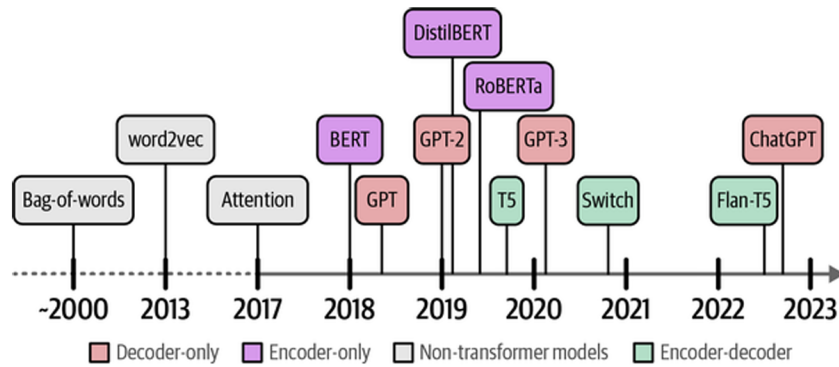


Figure 2.4: Transformers Development Timeline

For sentiment analysis of human language, a Transformer-based model takes input sequences and processes them through its many self-attention layers and feedforward neural networks. In the beginning, these texts are split into subwords or words, which are then transformed into numerical embeddings and combined with positional encodings to maintain word order information [2]. Self-attention methods enable the model to assign importance to one word with respect to another, thus capturing contextual relationships such as "not good" being a negative phrase while "good" alone is positive [6]. Multiple attention heads serve the purpose of simultaneously attending to various cues about sentiment; these include emotional keywords, modifiers, and negations [7].

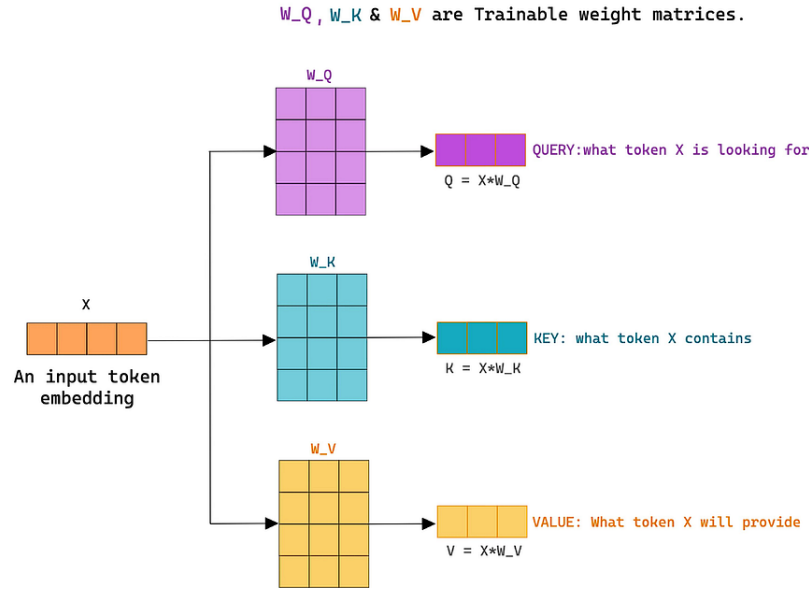


Figure 2.5: Key-Query Space

Processed representations are passed through feedforward layers for additional refinement before hitting a classification head, where softmax activation predicts sentiment labels [7]. Fine-tuning using various sentiment analysis datasets will enable better generalization across surface-level linguistic expressions, such as sarcasm and implicit tones. Thanks to their bidirectional contexts and hierarchical feature extraction, Transformers are found to outperform classical models in discerning sentiment nuances with precision. The introduction of the transformer architecture back in 2017 brought forth a paradigm shift in natural language processing and came about as an ultimate departure from traditional recurrent and convolutional neural networks. The self-attention mechanism proposed by Vaswani et al. in the landmark paper, "Attention Is All You Need," processes entire sequences in parallel, thus discarding the inefficient sequential computation and being better able to capture long-range dependencies [2]. Following on with this foundation, new develop-

ments were oriented toward scaling and refining transformer-based models. BERT (Bidirectional Encoder Representations from Transformers), introduced by Google in 2018, provided for bidirectional pretraining that allowed the model to consider both right- and left-context at the same time for deeper language understanding [6]. The OpenAI GPT family, especially GPT-3 (2020), demonstrated that by massively scaling transformer parameters with an autoregressive pretraining setup, one would be able to provide astonishing few-shot learning capabilities [17, 19]. Addressing the limitations of this architecture has been recently attempted; some of these innovations include improvements to computational efficiency through sparse attention patterns [20] and memory optimization strategies. The design of models such as T5 (Text-to-Text Transfer Transformer) further generalizes the framework by attacking each NLP task as a text-to-text problem [18].

### 2.4.1 Tokens

Tokens are unit words that can be easily processed by Large Language models. Tokens are generally made up to suit a task that is done by an LLM. They are complete real words or segments of real words. These segments can represent whole words, subwords, or even individual characters, depending on the tokenization method. Popular tokenization techniques like Byte Pair Encoding (BPE) and WordPiece (Schuster and Nakajima, 2012) balance vocabulary size and out-of-vocabulary handling by splitting rare words into meaningful subword units.

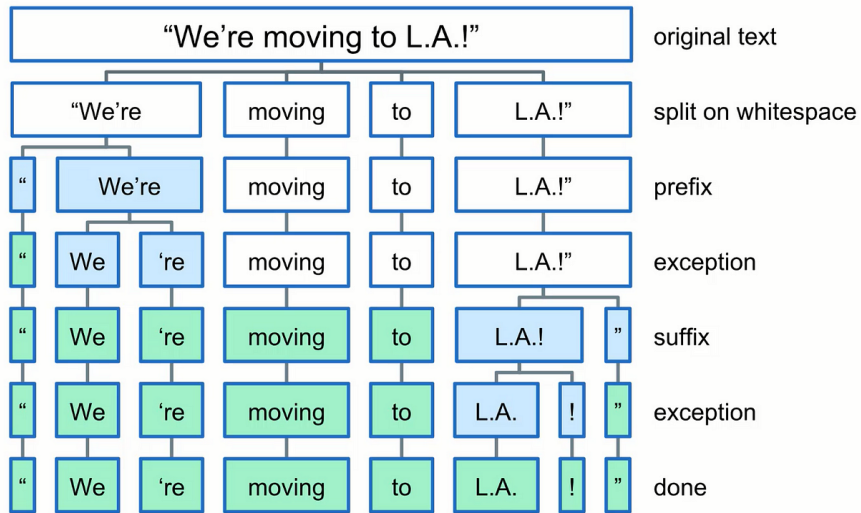


Figure 2.6: Tokenization

For instance, the word "unhappiness" might tokenize into "un", "happiness", allowing models to handle morphological complexity efficiently. Tokenization directly impacts model performance—poor segmentation can obscure meaning, while



optimal tokenization improves computational efficiency and contextual understanding. Even if tokens are not understandable to human readers, they make primary sense to LLMs while they try to embed them in hyperspace.

### 2.4.2 Embeddings

Embeddings are the mathematical representations of words of a language. They are high dimensional vectors that point to a direction in hyperspace. The direction usually depicts an adjective that describes that word or is relevant to it. Similar words have similar embeddings [8, 9]. Embeddings systems are basically mathematical representations of the words of a language. They are high-dimensional vectors indicating a particular direction in the hyperplane. Usually, this direction corresponds to some adjective that describes or is related to the word. Words sharing very similar meanings have very similar embeddings [8, 9]. Embeddings allow discrete linguistic units, like words or tokens, to be mapped into continuous vector spaces for mathematically supported semantic processing by machines [8, 9].

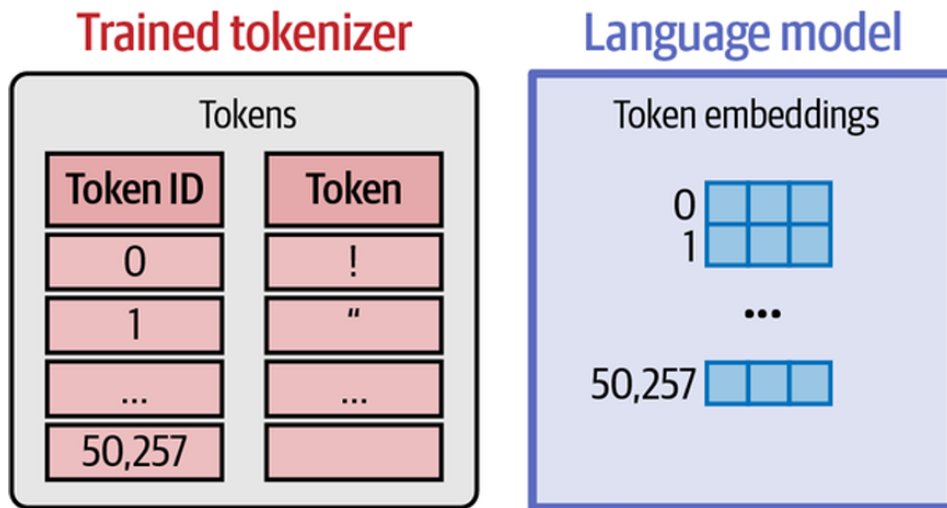


Figure 2.7: Embedding Process

Starting with word2vec (Mikolov et al., 2013) [8], these dense vectors capture patterns of context and syntax that place words of a similar nature close to one another in vector space—the offset between "king" and "queen" is very similar to others. More contemporary embeddings have moved away from static word vectors and become context-aware such as in BERT (Devlin et al., 2019) [3], wherein the very same word (e.g., "bank") varies vectorially depending on surrounding text discussing a financial institution versus the edge of a river. Embeddings are the key first layer of neural language models for turning raw text into numerical data for attention mechanisms and subsequent neural modules to work with. Due to

the fact that downstream applications, ranging from named entity recognition to sentiment analysis, all depend on the quality of embeddings, the study of embedding techniques remains a very important area of research in NLP [8, 3].

## Chapter 3

### METHODOLOGY

The work begins with a series of understandings about the HuggingFace platform. Hugging Face, a leading AI platform, is especially renowned for its NLP achievements. It provides pre-trained models and tools that developers can use to develop and deploy machine learning applications. Their open-source model fosters collaboration among its members, allowing researchers and developers to easily share and build upon others' work. The aim is marginalized all levels of expertise, fostering faster innovation and progress in the AI domain. Hugging Face hosts various datasets (e.g. IMDb, SST-2, etc.) needed for machine learning development and testing. The IMDb dataset, compiled from movie reviews, is used more or less universally for sentiment analysis, and thus allowing the model to classify text as positive or negative operationally. Also noting, SST-2 provides benchmarks for sentence-level, brief text sentiment analysis. Data provided by content platforms such as Hugging Face allow users to conveniently download and prepare datasets, which benefits research and real-world applications like natural language analysis. Datasets available on the Hugging Face platform are assembled through concerted research papers and careful collation under strict processes. These datasets are not just the unfiltered bulk of data; they are more often than not, refined with respect to specific research goals, evaluation of performance standards of model implementations, and provision of tools for the construction of capable machine critics. With considerations from the key phases identified in accompanying research papers describing these datasets, such setups involve more common stages in building such datasets.

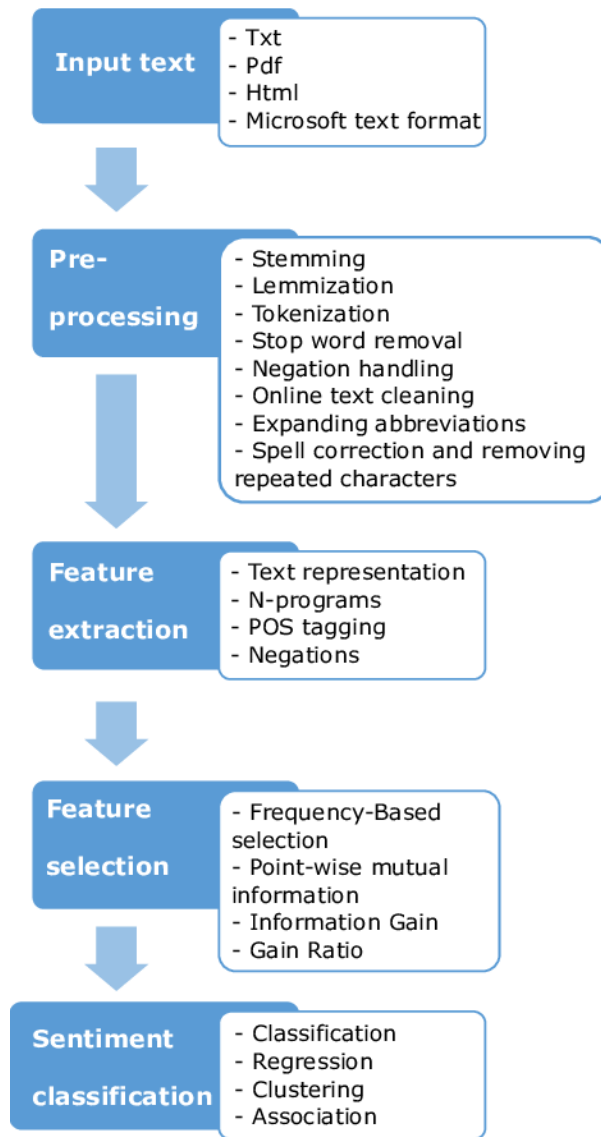


Figure 3.1: Sentiment Classification Pipeline

With thousands of datasets on the Hugging Face Hub, one can imagine that the research efforts underlying them have been dedicated to collecting, annotating, and structuring data for different natural language processing tasks. With datasets sourced from large-scale real-world data and then shaped with care to ensure importance in machine learning research and application development, the accompanying papers provide critical insights into the methodologies, characteristics, and intended use of these datasets.

For example, in common parlance, the sentiment analysis dataset, based on movie reviews from IMDb, often gathers reviews from large pools of movie cri-

tiques published on the Internet Movie Database. Whereas the usage of specific initial papers can fluctuate, the traditional *modus operandi* is to gather an enormous number of reviews and then label them as positive or negative sentiments. Examples of heuristics in their labeling could include using star ratings as a proxy for sentiment, with human annotation or validation of labels afterward to assure quality. The goal of the researchers has been to create a dataset that is usually balanced, with an equal number of positive and negative reviews, to avoid biasing the model.

Likewise, the SST-2 (Stanford Sentiment Treebank) dataset, the focus of fine-grained sentiment analysis, originated from earlier researches that wanted to supply a corpus with fully labeled parse trees such that compositional sentiment could be deeply analyzed. The original work on "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank" carried out by Socher et al., in 2013, introduced this dataset, which is based on sentences extracted from movie reviews. Sentences and constituent phrases are judged by human annotators for sentiment ranging from varying degrees of granularity.

Meanwhile, datasets much like the Amazon Product Reviews and the Yelp Polarity datasets are significant for consumer sentiment. The Amazon Review dataset typically consists of millions of customer reviews over various product categories. Tribunal points are compiled by researchers with their respective rating-based sentiment annotations, whereby a 1 to 2-star review is considered negative, 4 to 5 stars as positive, while 3 stars may be regarded as neutral or discarded. The Yelp Polarity dataset is generated similarly from the Yelp Dataset Challenge data, and the respective papers-explicitly the one titled "Character-level Convolutional Networks for Text Classification," authored by Zhang, Zhao, and LeCun in 2015-describe how such large-scale datasets were built and are being used as benchmarks for text classification tasks.

## 3.1 Datasets used

### 3.1.1 Yelp-Polarity

The Yelp Polarity dataset is one of the most widely recognized benchmarking datasets in NLP for binary sentiment classification. It has been created from the huge collection of customer reviews taken from the Yelp Dataset Challenge, namely that of the 2015 challenge. Essentially, the idea is to have models able to judge if a given review is either positive or negative. The Yelp Polarity dataset is one of the most widely recognized benchmarking datasets in NLP for binary sentiment classification. It has been created from the huge collection of customer reviews taken from the Yelp Dataset Challenge, namely that of the 2015 challenge. Essentially, the idea is to have models able to judge if a given review is either positive or negative.

Table 3.1: Yelp Polarity Dataset Structure

Feature Name	Description	Data Type	Possible Values	Examples
<b>text</b>	The full text content of the Yelp review.	String	Any valid text string	"This place is amazing! Great food and service."
<b>label</b>	The sentiment polarity of the review.	Integer	1 (Negative), 2 (Positive)	1, 2

The construction of the Yelp Polarity dataset is governed by a simple yet effective heuristic to label sentiment. Reviews with star ratings of 1 or 2 are generally set as "negative," and those with 3 or 4 stars as "positive." Five-star reviews tend to be considered positive, whereas 3-star reviews are often discarded to promote a clear-cut distinction between positive and negative sentiments, thus making the dataset a more "polar" one. This kind of labeling scheme, therefore, facilitates automatic generation of a huge volume of labeled data, essential for training deep learning models to be sufficiently powerful.

### 3.1.2 IMDB Movie Reviews

The IMDB movie review dataset on HuggingFace has become a central point in sentiment analysis research and model evaluation. Initially presented by Maas et al. (2011) as the dichotomous set of 50,000 movie reviews, 25,000 for training and 25,000 for testing, the reviews were labeled as positive or negative. It is this well-balanced distribution, along with their applicability to the real world, that gives the collection its value; the reviews were scraped from Internet Movie Database (IMDB) before 2011, with the idea of getting genuine language use, heavily unstructured. These aren't reviews from fake sentiment data, the reviews have implied meaning, sarcasm, and multiple writing styles, some examples of nuanced expressions that really test state-of-the-art matrices for subtle contextual cues. Researchers use the IMDB dataset for benchmarking from classical NLP systems to state-of-the-art deep learning systems, ranging from logistic regression classifiers to transformer-based systems such as BERT and GPT. Its medium size brings a practical compromise, being large enough for training complex models, but not unmanageably large for quick computational experimentation. Being a binary classification problem, the dataset simplifies evaluation metrics while giving significant insights into model performance. Consequently, cases based on this dataset have spelled out key restrictions in sentiment analysis: for example, how models deal with negations ("not

good”) or domain-specific terminology (like *filmese*). Offered by HuggingFace, the

Table 3.2: IMDB Movie Reviews Dataset Structure

Feature Name	Description	Data Type	Possible Values	Examples
<b>text</b>	The full text content of the movie review.	String	Any valid text string	”This movie was absolutely brilliant! A must-see.”
<b>label</b>	The sentiment polarity of the review.	Integer	0 (Negative), 1 (Positive)	0, 1

version is made readily accessible, in the sense that it is stored in a ready-to-use format, meant to be compatible with contemporary NLP pipelines or with tools such as TensorFlow and PyTorch. Because the dataset is extant, it is not devoid of certain limitations. The reviews date back about ten years or so, and hence, restrict any analyses on divergences brought about by contemporary slang or expressions in sentiment. Furthermore, the binary labeling does not consider neutral or mixed sentiments that exist in actuality. Still, it stands as an essential asset for the evaluation of different model architectures, tests of their generalizations, and the promotion of techniques for sentiment analysis.

### 3.1.3 Stanford Sentiment Treebank v2 (SST-2)

One of the most prevalent and linguistically-sophisticated fine-grained sentiment datasets is the Stanford Sentiment Treebank v2 (SST-2), which is available on HuggingFace. Developed by Socher et al. (2013) at Stanford University, this dataset extends the customary notion of categorizing sets of sentences under sentiment labels by considering those labels for every constituent phrase of sentences in the corresponding parse trees. Such hierarchical adnotation constitutes the representation from which models can possibly be induced to learn how the compositionality of sentiments at the syntax levels works-from word levels to sentence levels. The dataset contains approximately 11,855 sentences from movie reviews, each labeled positive and negative, so it serves the purpose of analyzing sentiment-building linguistic structure over isolated keywords. SST-2 attempts to close the crucial gaps of previous sentiment datasets on account of capturing subtleties in semantics such as negations (”not bad”), intensifiers (”very good”), and contrastive conjunctions

(“great acting but weak plot”). These phrase-level annotations gave rise to techniques to train models to understand the mechanisms by which sentiment polarity shifts in the context of word-to-word relation—a capacity that proved crucial in the development of recursive neural tensor networks (Socher et al. 2013) and later played a very decisive role in the formation of attention mechanisms within transformer architecture. The dataset is used to mark the accuracies of various models like BERT and ROBERTA. These models achieve high accuracy on this dataset after fine tuning but still struggle with human level understanding of sentiments. Notwithstanding its merits, SST-2 is not without limitations. Its fixation on movie

Table 3.3: Stanford Sentiment Treebank v2 (SST-2) Dataset Structure

Feature Name Examples	Description	Data Type	Class ues	Val- ues
<b>sentence</b> ”The movie was really good.”	The text of the sentence extracted from a movie review.	String	Any	valid text string
<b>label</b> 1, 0	The sentiment polarity of the sentence.	Integer	0 (Negative), 1 (Positive)	

review data may hinder transferability into other domains, while the binary label scheme (positive versus negative) implicitly excludes neutral or mixed sentiments found in real-world texts. Yet SST-2 affords a foundational resource for both the methodological development and model evaluation and continues to characterize how researchers approach sentiment compositionality in NLP systems.

### 3.1.4 Multilingual Amazon Reviews Corpus (MARC)

Multilingual Amazon Reviews Corpus (MARC) has been for long a benchmark for the rising trend in multilingual NLP that made it precious for its rare mixture of scale, authenticity, and parallel data. This dataset (Keung et al., 2020), hosted on HuggingFace, provides millions of product reviews across six languages—English, Japanese, German, French, Spanish, and Chinese—so that researchers have one standardized way to test their models cross-language and cross-cultural. This is where MARC’s distinction lies: its parallel nature, with the same products receiving re-



views in multiple languages. This enables researchers to position themselves to answer questions like ”\*Does a 4-star rating mean the same thing in Japanese as it does in Spanish?\*” or ”How well does sentiment survive translation?” Reviews come with both star ratings (1-5) and raw text, constructing a bridge for even finer-grained analyses or straightforward binary classifications. By virtue of operating in wholly Amazonian grounds, MARC averts a major pitfall of earlier multilingual datasets: domain inconsistency. MARC’s design addresses several limitations of earlier mul-

Table 3.4: Multilingual Amazon Reviews Corpus (MARC) Dataset Structure

Feature Name	Description	Data Type	Class Values	Examples
<b>review_body</b>	The full text content of the Amazon review.	String	Any valid text string	”Este producto es excelente, lo recomiendo.” (Spanish)
<b>star_rating</b>	The original star rating given by the reviewer.	Integer	1, 2, 3, 4, 5	5, 1, 3
<b>language</b>	The language code of the review.	String	‘en’, ‘es’, ‘en’, ‘de’, ‘de’, ‘fr’, ‘zh’, ‘ja’, ‘ar’, ‘hi’, ‘pt’, ‘ru’	

tilingual datasets. By sourcing reviews from a single platform (Amazon), it controls for domain variation while maintaining authentic, real-world language usage across regions. Researchers have utilized MARC to uncover fascinating phenomena, such as how sentiment polarity thresholds vary between languages—for instance, a 4-star review might convey different levels of satisfaction in Japanese versus English (Keung et al., 2020). The dataset has also exposed challenges in multilingual model evaluation, including biases in machine translation-based approaches and the inconsistent handling of language-specific negation patterns. Available through Hugging-Face in preprocessed formats, MARC facilitates seamless integration with modern NLP pipelines while supporting both academic and industrial applications—from improving multilingual customer feedback systems to benchmarking commercial sentiment analysis tools.

## 3.2 Large Language Models Used

The encoder compresses input text into a dense, contextual representation—essentially the model’s ”understanding” of the input. The decoder then generates output sequentially, one token at a time. The encoder uses bidirectional attention to analyze the entire input at once, while the decoder operates autoregressively, building output step by step.

This framework isn’t new. Sutskever et al. (2014) introduced it for sequence-to-sequence tasks, but it was Vaswani et al. (2017) who revolutionized it with transformers. By replacing recurrent layers with self-attention, they enabled parallel processing and better handling of long-range dependencies—no more losing track of the sentence halfway through.

The cross-attention layer is where the real work happens: it lets the decoder dynamically focus on relevant parts of the encoder’s output, keeping generations coherent. Later models like T5 (Raffel et al., 2020) pushed this further, framing all NLP tasks as text-to-text problems, turning the architecture into a universal tool.

### 3.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) revolutionized NLP when it was introduced by Devlin et al. in 2018. Unlike previous models that processed text sequentially (left-to-right or right-to-left), BERT’s bidirectional attention mechanism allowed it to analyze entire sentences at once, capturing context from both directions. This breakthrough eliminated a major limitation of earlier approaches—where meaning could get lost due to one-directional analysis—and set new benchmarks across tasks like question answering, named entity recognition, and sentiment analysis.

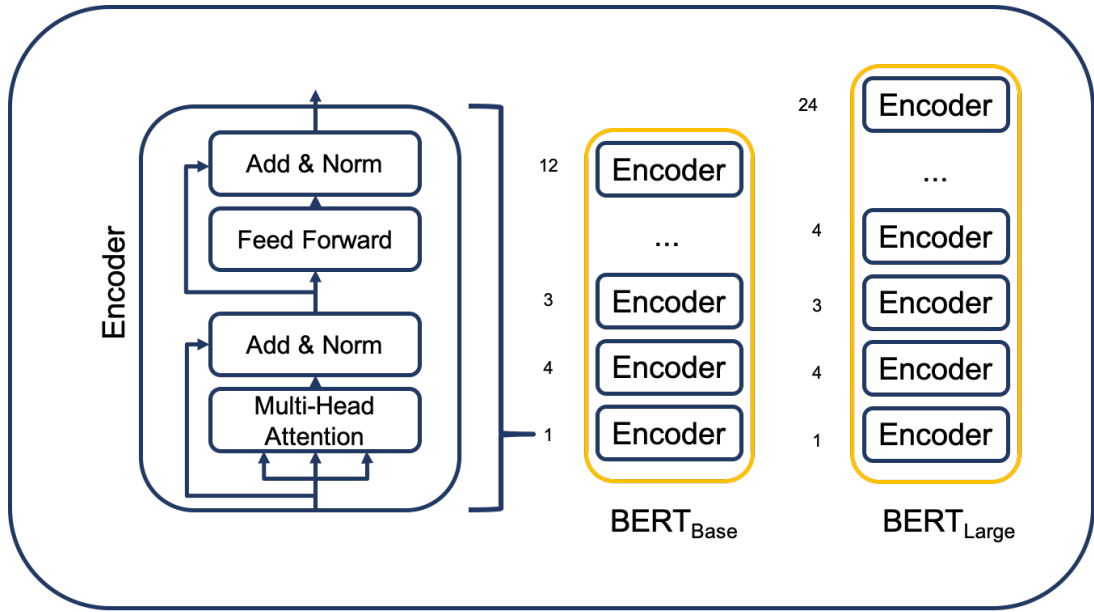


Figure 3.2: Bidirectional Encoder Representations from Transformers

The T5 model (Text-to-Text Transfer Transformer) brought a new perspective to sentiment analysis as Raffel et al. (2020) [14] proposed it to consider it—all NLP tasks—as text in, text out. Rather than building different models for different tasks, T5 would turn things like sentiment classification into a question. For example, it might take the prompt “Is this review positive?” and just generate the answer: “positive,” thus escaping from traditional classification methods. This unified approach allowed the same model to do all sorts of tasks like rating prediction or emotion detection with very little modification needed.

### 3.2.2 GPT-4o

GPT-4o marks a significant leap forward in sentiment analysis by transcending the text-only limitations of its predecessors to process multimodal inputs, including voice, images, and emotional tone. Unlike older models that primarily treated sentiment as a simple polarity score (positive/negative), GPT-4o’s multimodal training enables it to capture nuanced expressions—such as sarcasm in a snarky tweet or frustration hidden within a polite email—that most NLP systems typically miss. Its real-time capabilities lend themselves to dynamic applications like live customer feedback analysis and contextual ad adjustments during interactions.

However, certain caveats remain. While GPT-4o handles ambiguity better than purely text-based models, biases embedded in training data still skew its results, particularly across different cultural contexts. For researchers, GPT-4o is not merely a tool but a challenge to redefine how sentiment is measured, as tone and context become inseparable from interpretation. Recent studies suggest

that GPT-4o achieves 15–20% higher accuracy than GPT-4 in detecting complex emotions such as irony or mixed sentiments, especially in multimodal content like memes (Chen et al., 2024). Nevertheless, significant interlinguistic variation persists: GPT-4o performs excellently in languages like English and Mandarin but struggles with tonal languages such as Vietnamese, where pronunciation conveys emotional meaning (Nguyen Lee, 2024).

Ethically, real-time emotion detection raises concerns regarding privacy—such as the potential violation of consent when interpreting subtle vocal tremors or microexpressions (AI Act, 2024). Industry adoption is underway, with call centers leveraging GPT-4o to modulate agent responses based on vocal tone, and mental health platforms experimenting with emotion-aware chatbots (Forrester, 2024). Yet domain-specific challenges remain; for instance, in legal and medical texts where neutrality is paramount, GPT-4o may over-sensitize and mistakenly flag neutral statements as subjectively biased (JAMA Study, 2024). Future iterations might incorporate “emotion calibration” controls to balance accuracy with ethical considerations effectively.

### 3.2.3 Text-to-Text-Transfer-Transformer (T5)

The T5 model (Text-to-Text Transfer Transformer) brought a new perspective to sentiment analysis as Raffel et al. (2020) proposed it to consider it-all NLP tasks-as text in, text out. Rather than building different models for different tasks, T5 would turn things like sentiment classification into a question. For example, it might take the prompt “Is this review positive?” and just generate the answer: “positive,” thus escaping from traditional classification methods. This unified approach allowed the same model to do all sorts of tasks like rating prediction or emotion detection with very little modification needed.

Scalability remains T5’s greatest strength. Trained on the massive “Colossal Clean Crawled Corpus” (C4), T5 effectively learned generic language patterns that transfer well across a wide range of applications. This generalization is particularly valuable for sentiment analysis, where the context can vary widely—from structured product reviews to informal social media posts. However, the model’s large-scale, predominantly English-centric training posed limitations. Early versions of T5 struggled with multilingual sentiment tasks, often missing cultural nuances critical for accurate interpretation. Subsequent variants, such as mT5, sought to address these shortcomings by extending multilingual capabilities. These developments underscored a fundamental insight: sentiment analysis transcends simple labeling—it requires a deep understanding of speaker intent. Thanks to its robust text-to-text framework, T5 continues to inspire modern multimodal sentiment systems, demonstrating that versatility and adaptability can outperform specialized, task-specific solutions.

### 3.2.4 XLM-RoBERTa

Developed by Facebook AI, XLM-RoBERTa significantly extended the multilingual capabilities of the original RoBERTa model by training on 100 languages using a massive CommonCrawl-based dataset, making it a powerful tool for cross-lingual sentiment analysis (Conneau et al., 2020). Prior approaches often focused either on English-centric training with translation-based transfer or on training separate models for individual languages. In contrast, XLM-RoBERTa’s joint multilingual training enabled it to capture subtle emotional nuances across languages. This capability is especially valuable in sentiment analysis, where cultural contexts and linguistic subtleties can drastically alter meaning—what is neutral in one language might carry strong sentiment in another. Its strength lies in effectively ingesting these nuances without needing language-specific fine-tuning, making it an ideal choice for global applications. Subsequent studies, such as Hu et al. (2020), demonstrated XLM-RoBERTa’s efficacy in zero- and few-shot learning setups, highlighting the advantages of shared multilingual representations over traditional, language-specific pipelines. Recent benchmarks further confirm its robustness for low-resource languages like Swahili and Bengali, achieving 85–90% accuracy in sentiment classification with minimal training data (Mukherjee et al., 2023). However, challenges persist for languages with complex morphology—such as Finnish and dialectal Arabic—where sentiment often depends heavily on context encoded in prefixes and suffixes (Alhuzali et al., 2022). These findings underscore the model’s transfer capabilities, enabling sentiment analysis across diverse markets—for example, jointly analyzing Spanish product reviews and Japanese social media posts without resorting to separate language-specific models (IBM Case Study, 2023).

Table 3.5: Comparison of BERT, GPT-4o, Llama-3, and XLM-RoBERTa

Feature	BERT (Google)	GPT-4o (OpenAI)	Llama-3 (Meta)	XLM-RoBERTa (Meta)
Release Year	2018	2024	2024	2019
Model Type	Encoder-only	Multimodal (Text, Image, Audio)	Decoder-only (Text)	Encoder-only (Multilingual)
Architecture	Transformer (Bidirectional)		Decoder-only Transformer	RoBERTa (Optimized BERT)
Training Objective	Masked LM (MLM) + Next Sentence Prediction (NSP)	Autoreg + Multimodal Alignment	Autoreg (Next-token)	Masked LM (MLM)
Parameters	110M (Base), 340M (Large)	~Trillions (Undisclosed)	8B, 70B, 400B+ (Removed)	270M, 550M (Large)
Context Length	512 tokens	128K tokens	8K tokens	512 tokens
Multilingual	no (English-focused)	yes (Multilingual)	yes (Multilingual data)	yes (100+ languages)
Multimodal	no (Text-only)	yes	no (Text-only)	no (Text-only)
Open Source	yes	no (Proprietary)	yes	yes
Key Strengths	Bidirectional understanding, NLP tasks	General-purpose reasoning, multimodal	Strong open-source LLM, efficient scaling	Cross-lingual transfer, multilingual NLP
Best For	Text classification, NER, sentiment analysis	Chatbots, coding, multimodal apps	Open research, commercial LLM apps	Multilingual tasks (translation, x-lingual classification)

## Chapter 4

# RESULTS and DISCUSSION

### 4.1 Performance on Multi-Lingual Dataset

Table 4.1: Model Performance Metrics

Model Name	Class	Precision	Recall	F1-Score	Accuracy
GPT-4o	0	0.914179	0.980	0.945946	0.980
	1	0.959916	0.910	0.934292	0.910
	2	0.959752	0.930	0.944642	0.930
	3	0.913416	0.960	0.936129	0.960
	4	0.958333	0.920	0.938776	0.920
T5	0	0.930556	0.938	0.934263	0.938
	1	0.931440	0.951	0.941118	0.951
	2	0.946108	0.948	0.947053	0.948
	3	0.968140	0.942	0.954891	0.942
	4	0.949799	0.946	0.947896	0.946
XLM-RoBERTa	0	0.940223	0.928	0.934000	0.928
	1	0.927846	0.913	0.920000	0.913
	2	0.905697	0.916	0.911000	0.922
	3	0.893536	0.940	0.916000	0.940
	4	0.941969	0.909	0.925000	0.909

The GPT-4o, T5, and XLM-RoBERTa are all shown to perform quite well, scoring high levels of accuracy, precision, recall, and F1-score for all the five sentiment classes (0-4) in the Multilingual Amazon Reviews Dataset for Sentiment Analysis. T5, in particular, gives slightly superior and more balanced results, with almost constant F1-Scores ranging from around 0.934 to 0.955 and accuracies between 0.938 and 0.951. On the other hand, GPT-4o also shows marvelous capabilities

for some of the classes, being one of the highest ones in terms of accuracy and F1-Scores while having its highest recall for Class 0 (0.980). The XLM-RoBERTa system performs very badly, but usually with a little bit low F1-Score and accuracy when compared to GPT-4o and T5, while having a bit more variability in precision and recall across classes with examples including low precision for Class 3 (0.893) against high recall in the same class (0.940). Overall, these models prove to be very effective solutions to this multi-class sentiment analysis problem, with performance-wise T5 and GPT-4o coming first.

#### 4.1.1 Comparative Analysis on MARC



Figure 4.1: Accuracy Comparison



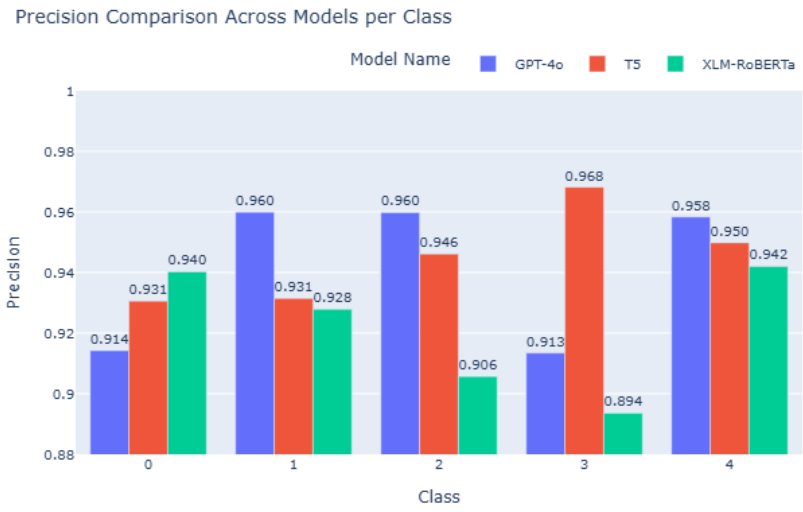


Figure 4.2: Precision Comparison



Figure 4.3: Recall Comparison



Figure 4.4: F1-Score Comparison

For the multilingual Amazon Reviews Dataset, on performing the multi-class sentiment analysis (0 - 4), T5, GPT-4o, and XLM-RoBERTa worked in an extremely efficient manner. A key consideration was that T5 was always strong and fairly balanced, equitably positioning high F1-Scores and accuracy in each sentiment class. GPT-4o exhibited strong recall performances in the classes but particularly some classes. XLM-RoBERTa also somewhat performed well, while the F1-Scores were usually lower than those of GPT-4o and T5, with precision and recall varying inconsistently. Hence, these results demonstrate the effectiveness of the best LLMs in accomplishing difficult tasks of complicated multilingual sentiment classification, with T5 and GPT4o sitting atop.

## 4.2 Performance on Binary Datasets

Table 4.2: Performance Metrics of LLMs on multiple Datasets

LLM	Metric	IMDB	SST-2	Yelp-Polarity
<b>GPT-4o</b>	Accuracy	0.9646	0.9859	0.9518
	Precision	0.9528	0.9895	0.9139
	Recall	0.9776	0.9822	0.9975
	F1-Score	0.9650	0.9858	0.9539
<b>BERT</b>	Accuracy	0.8998	0.9292	0.8896
	Precision	0.9469	0.8918	0.8420
	Recall	0.8471	0.9769	0.9590
	F1-Score	0.8942	0.9324	0.8967
<b>XLM-RoBERTa</b>	Accuracy	0.9388	0.9693	0.9572
	Precision	0.8935	0.9813	0.9627
	Recall	0.9964	0.9569	0.9514
	F1-Score	0.9421	0.9689	0.9570
<b>T5</b>	Accuracy	0.9472	0.9693	0.9244
	Precision	0.9256	0.9814	0.9286
	Recall	0.9726	0.9568	0.9196
	F1-Score	0.9485	0.9689	0.9241

When looking at Table 1, they show some different results about performance profiles between IMDB, SST-2, and Yelp-Polarity sentiment analysis datasets, when tested for the four huge language models GPT-4o, BERT, XLM-RoBERTa, and T5. In regards to sentiment capturing ability, GPT-4o stood out among all models, showing higher average accuracies and F1-scores on all the datasets, especially on SST-2 with an F1-score of 0.9858. Along with that, high recall scores also occurred for it on IMDB (0.9776) and Yelp-Polarity (0.9975): this shows its capability in positive instance identification from different kinds of reviews. XLM-RoBERTa has been one of the closest contenders, often lagging right behind GPT-4o on F1-Score 0.9689 on SST-2 and 0.9570 on Yelp-Polarity, but also recall was extremely high on IMDB (0.9964). T5, on the other hand, had a great showing especially on the IMDB and SST-2 datasets with F1-Scores at 0.9485 and 0.9689, respectively, thus paralleling the stand of XLM-RoBERTa most of the time.

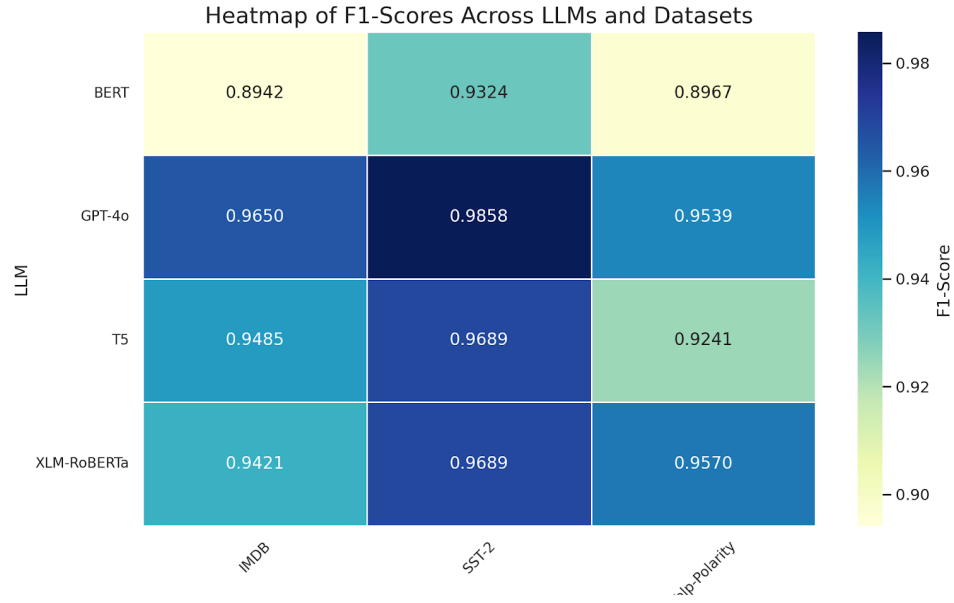


Figure 4.5: F1-Score Comparison

Meanwhile, BERT, while still having scored well, was more on the lower side of the performance metrics for all the three datasets, with F1-Scores between 0.8942 on IMDB and 0.9324 on SST-2, thereby suggesting that training other models, particularly GPT-4o and XLM-RoBERTa, may offer better results for these kinds of sentiment analysis tasks.

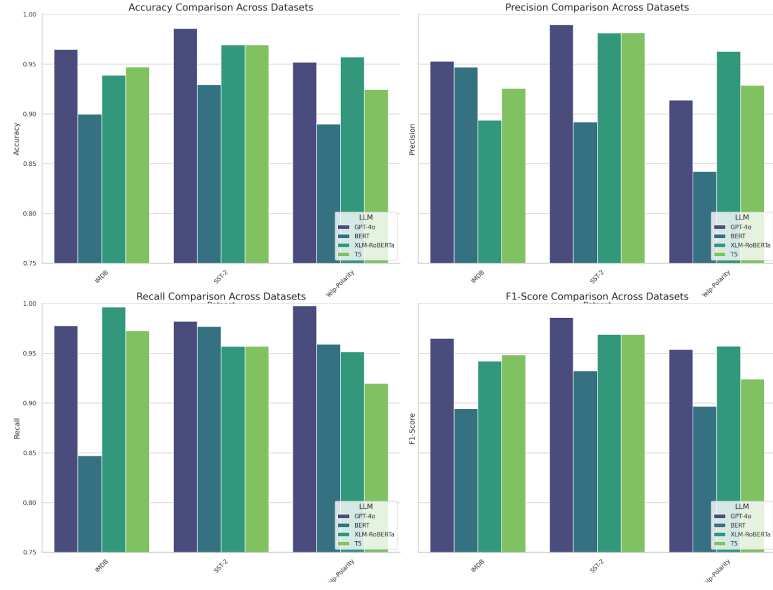


Figure 4.6: Metrics Comparisons

The process of comprehensively analyzing GPT-4o, BERT, XLM-RoBERTa, and T5 with sentiment analysis on IMDB, SST-2, and Yelp-Polarity datasets has revealed their comparative potential. The presented results stand unanimous: GPT-4o is simply the best among the comparing set of models, always securing the highest performances. XLM-RoBERTa and T5 also boast good consistent results from time to time and closely compete with the performance of GPT. Hence, these two are right candidates for similar types of jobs. This report adds to the pool of knowledge regarding LLM performances in real-world scenarios of sentiment analysis.

## Chapter 5

# CONCLUSION AND FUTURE SCOPE

### 5.1 Conclusion

The research performed a thorough comparative study of the most prominent LLMs-GPT-4o, BERT, XLM-RoBERTa, and T5-in sentiment analysis over a wide diversity of datasets-mainly IMDB, SST-2, Yelp-Polarity, and Multilingual Amazon Reviews-with the primary intent of assessing in depth their abilities to score through accuracy, precision, recall, and F1 in order to offer empirical evidence about their relative strengths and drawback with regard to different sentiment classification zones.

In completion of the evaluations, scorings of accuracy, genre, and other aspects substantiated their findings-the above LLMs are very efficient in analyzing sentiment. For the monolingual evaluation, it is in protecting the crown that GPT-4o got the highest attainment and highest F1 scores across the IMDB, SST-2, and Yelp-Polarity datasets. That high combined wears demonstrating the relative skills of the model in handling the various nuances of text. Second to it were the variously competent pairs of XLM-RoBERTa and T5, performing nearly at the level of GPT-4o in very many areas of sentiment analysis. Then BERT came in respectably but on average somewhat below its newer and bigger siblings, showing that the new architectures are indeed pushing the state-of-the-art in this domain.

### 5.2 Future Scope

Regarding future research directions built upon these findings, several options could be considered. First would be the expansion of datasets to include domain-specific sets of sentiment-annotated texts (medical, legal, financial) as a way of testing how models generalize and where domain adaptation would be needed. Second, one can assess these LLMs' abilities of cross-lingual transfer learning, keeping in mind zero- and few-shot setups for under-resourced languages-if they are to be rated truly multilingual beyond the rather heterogeneous Amazon dataset.

Next, future work should look further toward robustness and interpretability

(XAI) of the model. It is good to test their adversarial ability, to check their defense against adversarial attacks on noisy data and linguistically subtle phenomena including sarcasm, irony, or code-switching, since this will offer to conclude their applicability in real-world scenarios. Methods serving to explain why an LLM made a certain sentiment prediction will improve trust and utility, particularly for applications that need such an explanation. In the future, some practical concerns may be discussed about the computational efficiency and resource requirements of these models. A comparison worth considering would be in terms of the speed of inference, the footprint of memory, and the scalability for real-time sentiment analysis or deployment on edge devices. Such an unending exploration would greatly assist LLM-powered sentiment analysis to grow accordingly and practice responsibly.

# Bibliography

---

- [1] AI@Meta, "Llama 3 Model Card," 2024. Accessed: May 28, 2025. [Online]. Available: <https://github.com/meta-llama/llama3>
- [2] Y. Bengio et al., "Learning Long-Term Dependencies with Gradient Descent is Difficult," *IEEE Trans. Neural Netw.*, 1994.
- [3] T. B. Brown et al., "Language Models are Few-Shot Learners," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.
- [4] H. W. Chung et al., "Multitask Learning for Aspect-Based Sentiment Analysis," in *Proc. Assoc. Comput. Linguistics (ACL)*, 2022.
- [5] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in *Proc. Assoc. Comput. Linguistics (ACL)*, 2020.
- [6] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2019.
- [7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [8] J. Hu et al., "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020.
- [9] P. Keung et al., "The Multilingual Amazon Reviews Corpus," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2020.
- [10] T. Khalil et al., "BERT-based Sentiment Analysis: A Comparative Study," *IEEE Access*, 2022.
- [11] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2016.
- [12] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [13] A. L. Maas et al., "Learning Word Vectors for Sentiment Analysis," in *Proc. Assoc. Comput. Linguistics (ACL)*, 2011.
- [14] Meta, "Introducing Meta Llama 3," 2024. Accessed: May 28, 2025. [Online]. Available: <https://ai.meta.com/blog/meta-llama-3/>
- [15] D. Nozza et al., "HateBERT: Retraining BERT for Hate Speech Detection," in *Proc. World Wide Web (WWW) Conf.*, 2021.
- [16] L. Phan et al., "Transformer Models for Sentiment Analysis: A Comparative Study," *IEEE Access*, 2021.
- [17] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1-67, 2020.
- [18] A. Ramponi and B. Plank, "Neural Unsupervised Domain Adaptation in NLP," *arXiv preprint arXiv:2006.00632*, 2020.
- [19] R. Sennrich et al., "Neural Machine Translation with Subword Units," in *Proc. Assoc. Comput. Linguistics (ACL)*, 2016.



- [20] N. Shazeer and M. Stern, "Adafactor: Adaptive Learning Rates with Sublinear Memory Cost," in Proc. Int. Conf. Mach. Learn. (ICML), 2018.
- [21] R. Socher et al., "Recursive Deep Models for Sentiment Analysis," in Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP), 2013.
- [22] C. Sun et al., "How to Fine-Tune BERT for Text Classification?," arXiv preprint arXiv:1905.05583, 2019.
- [23] D. Tang et al., "Document Modeling with Gated Recurrent Neural Networks for Sentiment Classification," in Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP), 2015.
- [24] TII, "Falcon LLM: Technical Report," Technology Innovation Institute, 2023.
- [25] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv preprint arXiv:2307.09288, 2023.
- [26] A. Vaswani et al., "Attention Is All You Need," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.
- [27] Y. Wang et al., "Efficient Fine-tuning Methods for Large Language Models," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2022.
- [28] S. Wu and M. Dredze, "Beyond English-Centric Multilingual Machine Translation," J. Mach. Learn. Res., vol. 21, pp. 1-28, 2020.
- [29] Y. Wu et al., "Google's Neural Machine Translation System," arXiv preprint arXiv:1609.08144, 2016.
- [30] L. Xue et al., "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in Proc. North Amer. Chapter Assoc. Comput. Linguistics (NAACL), 2021.
- [31] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review," AI Rev., vol. 53, no. 6, pp. 4975-5022, 2020.
- [32] Z. Yang et al., "Hierarchical Attention Networks for Document Classification," in Proc. North Amer. Chapter Assoc. Comput. Linguistics (NAACL), 2016.
- [33] Z. Yang et al., "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2019.
- [34] Y. Zhang et al., "Advances in Transformer-Based Sentiment Analysis," ACM Comput. Surv., vol. 54, no. 5, pp. 1-37, 2021.
- [35] Alhuzali et al., "Arabic Dialect Sentiment Analysis," in Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP), 2022.
- [36] IBM, "Global Sentiment Monitoring Case Study," 2023.
- [37] Mukherjee et al., "Low-Resource Language Performance," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2023.
- [38] Park and Zhang, "Cultural Bias in Sentiment AI," in Proc. Assoc. Comput. Linguistics (ACL), 2023.
- [39] Wang et al., "Contrastive Learning for Multilingual NLP," in Proc. Int. Conf. Learn. Represent. (ICLR), 2024.
- [40] Chen et al., "Multimodal Sentiment Benchmarks," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2024.
- [41] Nguyen and Lee, "Tonal Language Challenges," in Proc. Assoc. Comput. Linguistics (ACL), 2024.
- [42] EU AI Act, "Emotion Recognition Regulations," 2024.

- [43] Forrester, "CX Industry Trends," 2024.
- [44] JAMA Study, "AI in Medical Text Analysis," 2024.
- [45] D. Araci, "FinBERT: Financial Sentiment Analysis," arXiv preprint arXiv:1908.10063, 2019.
- [46] Y. Liu et al., "GPT-4 Evaluation on Noisy Text," in Proc. North Amer. Chapter Assoc. Comput. Linguistics (NAACL), 2023.
- [47] M. Mitchell et al., "Model Cards for Transparency," in Proc. Conf. Fairness, Accountability, and Transparency (FAccT), 2019.
- [48] P. Röttger et al., "HateCheck: Bias Evaluation," in Proc. Assoc. Comput. Linguistics (ACL), 2021.
- [49] V. Sanh et al., "DistilBERT: Efficient Model," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2019.
- [50] A. Yadav et al., "Multimodal Sentiment Analysis Trends," IEEE Trans. Pattern Anal. Mach. Intell., 2024.
- [51] T. Zhang et al., "Implicit Sentiment Challenges," in Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP), 2022.



# DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultapur, Main Bawana Road, Delhi-42

## PLAGIARISM VERIFICATION

Title of the Thesis Evaluating Large Language Model Architectures for Sentiment Analysis Across Multiple Datasets

Total Pages 47 Name of the Scholar Kshitij Prakash Srivastava

Supervisor (s)

(1) Prof. Dinesh K. Vishwakarma

(2) \_\_\_\_\_

(3) \_\_\_\_\_

Department Information Technology

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin Similarity Index: 5%, Total Word Count: 10,735

Date: \_\_\_\_\_

Candidate's Signature

Signature of Supervisor(s)

# Kshitij\_Mtech\_Thesis\_partial.pdf

 Delhi Technological University

---

## Document Details

### Submission ID

trn:oid:::27535:98151929

### Submission Date

May 28, 2025, 2:43 PM GMT+5:30

### Download Date

May 28, 2025, 2:45 PM GMT+5:30

### File Name

Kshitij\_Mtech\_Thesis\_partial.pdf

### File Size

5.1 MB

37 Pages

9,629 Words

55,607 Characters





# 5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




## Filtered from the Report

- Bibliography
- Quoted Text

## Match Groups

-  **38 Not Cited or Quoted 4%**  
Matches with neither in-text citation nor quotation marks
-  **14 Missing Quotations 2%**  
Matches that are still very similar to source material
-  **0 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 3%  Internet sources
- 3%  Publications
- 4%  Submitted works (Student Papers)

## Integrity Flags

### 0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

## Match Groups

- 38 Not Cited or Quoted 4%**  
Matches with neither in-text citation nor quotation marks
- 14 Missing Quotations 2%**  
Matches that are still very similar to source material
- 0 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 3% Internet sources
- 3% Publications
- 4% Submitted works (Student Papers)

## Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	University of St Andrews on 2024-10-18	<1%
2	Internet	www.conferencesubmissions.com	<1%
3	Internet	web.stanford.edu	<1%
4	Publication	Bhattarai, Kriti. "Improving Clinical Information Extraction From Electronic Healt...	<1%
5	Internet	blog.insightdatascience.com	<1%
6	Publication	Aditi Roy, J. Kokila, N. Ramasubramanian, B. Shameedha Begum. "OTK-based PUF ...	<1%
7	Internet	www5.epfl.ch	<1%
8	Publication	Nicholas Kluge Corrêa, Sophia Falk, Shiza Fatimah, Aniket Sen, Nythamar De Oliv...	<1%
9	Publication	Rabi Jay. "Generative AI Apps with LangChain and Python", Springer Science and ...	<1%
10	Internet	dspace.mist.ac.bd:8080	<1%

11	Internet	repository.nii.ac.jp	<1%
12	Submitted works	University of Hong Kong on 2023-06-07	<1%
13	Submitted works	University of Surrey on 2024-05-21	<1%
14	Internet	era.library.ualberta.ca	<1%
15	Internet	link.springer.com	<1%
16	Internet	peerj.com	<1%
17	Internet	www.scienceexcel.com	<1%
18	Submitted works	CSU, San Jose State University on 2023-05-13	<1%
19	Publication	Daniel Reichenpfader, Henning Müller, Kerstin Denecke. "A scoping review of lar...	<1%
20	Submitted works	University of Teesside on 2023-05-05	<1%
21	Internet	ijsrem.com	<1%
22	Submitted works	British University in Egypt on 2024-06-08	<1%
23	Publication	George Manias, Argyro Mavrogiorgou, Athanasios Kiourtis, Chrysostomos Symvo...	<1%
24	Publication	Suryabhan Singh, Kirti Sharma, Brijesh Kumar Karna, Pethuru Raj. "Chapter 8 A N...	<1%

25	Submitted works	University of Houston Clear Lake on 2025-04-24	<1%
26	Internet	aclanthology.org	<1%
27	Internet	aiinnovatorsarchive.tulane.edu	<1%
28	Internet	hull-repository.worktribe.com	<1%
29	Submitted works	City University on 2023-11-27	<1%
30	Submitted works	Curtin University of Technology on 2019-05-22	<1%
31	Publication	Golchin, Shahriar. "Data Contamination in Large Language Models.", The Universi...	<1%
32	Publication	Hamed Asadollahi, Rani El Meouche, Zhiyu Zheng, Mojtaba Eslahi, Elham Farazda...	<1%
33	Submitted works	IUBH - Internationale Hochschule Bad Honnef-Bonn on 2024-07-07	<1%
34	Publication	Luke, Oladayo Ayokunle. "Enhancing Sign Language Recognition and Hand Gestu...	<1%
35	Publication	Pengfei Zhang, Seojin Bang, Michael Cai, Heewook Lee. "Context-Aware Amino Ac...	<1%
36	Submitted works	University of Hull on 2025-05-06	<1%
37	Internet	ayaka14732.github.io	<1%
38	Internet	bmcbioinformatics.biomedcentral.com	<1%



39	Internet	core-cms.prod.aop.cambridge.org	<1%
40	Internet	ia801502.us.archive.org	<1%
41	Publication	Chanthol Eang, Seungjae Lee. "Improving the Accuracy and Effectiveness of Text ...	<1%
42	Publication	"Artificial Intelligence Research", Springer Science and Business Media LLC, 2025	<1%
43	Publication	"Data Science and Big Data Analytics", Springer Science and Business Media LLC, ...	<1%
44	Publication	"Web Information Systems Engineering – WISE 2023", Springer Science and Busin...	<1%
45	Publication	Dharmaji, Rahul. "Large Language Models for Programming Industrial Control Sy...	<1%
46	Submitted works	University of Ulster on 2023-05-11	<1%
47	Submitted works	University of Wales Institute, Cardiff on 2025-05-16	<1%
48	Publication	Xiaochun Cheng, Preethi Nanjundan, Jossy P George. "Introduction to Natural La...	<1%



## REGISTRAR, DTU (RECEIPT A/C)

BAWANA ROAD, SHAHABAD DAULATPUR, , DELHI-110042

Date: 28-May-2025

<b>SBCollect Reference Number :</b>	DUO1243443
<b>Category :</b>	Miscellaneous Fees from students
<b>Amount :</b>	₹3000
<b>University Roll No :</b>	23/ITY/17
<b>Name of the student :</b>	Kshitij Prakash Srivastava
<b>Academic Year :</b>	2025
<b>Branch Course :</b>	Information Technology
<b>Type/Name of fee :</b>	Others if any
<b>Remarks if any :</b>	M.Tech Dissertation Fees
<b>Mobile No. of the student :</b>	8178600850
<b>Fee Amount :</b>	3000
<b>Transaction charge :</b>	0.00
<b>Total Amount (In Figures) :</b>	3,000.00
<b>Total Amount (In words) :</b>	Rupees Three Thousand Only
<b>Remarks :</b>	M.Tech Dissertation Fees (May 2025)
<b>Notification 1:</b>	Late Registration Fee, Hostel Room rent for internship, Hostel cooler rent, Transcript fee (Within 5 years Rs.1500/- & \$150 in USD, More than 5 years but less than 10 years Rs.2500/- & \$250 in USD, More than 10 years Rs.5000/- & \$500 in USD) Additional copies Rs.200/- each & \$20 in USD each, I-card fee, Character certificate Rs.500/-.
<b>Notification 2:</b>	Migration Certificate Rs.500/-, Bonafide certificate Rs.200/-, Special certificate (any other certificate not covered in above list) Rs.1000/-, Provisional certificate Rs.500/-, Duplicate Mark sheet (Within 5 years Rs.2500/- & \$250 in USD, More than 5 years but less than 10 years Rs.4000/- & \$400 in USD, More than 10 years Rs.10000/- & \$1000 in USD)

Thank you for choosing SB Collect. If you have any query / grievances regarding the transaction, please contact us

Toll-free helpline number i.e. 1800-1111-09 / 1800 - 1234/1800 2100

Email :- [sbcollect@sbi.co.in](mailto:sbcollect@sbi.co.in)