BERT vs DistilBERT: Enhanced Sentiment Analysis with Neutral Detection and Sarcasm Handling

Srujana Duggineni
MSc in Data Science
South East Technological University, Carlow, Ireland
C00313483@setu.ie

ABSTRACT

This paper offers a robust comparative assessment of transformer-based architectures for sentiment analysis and investigates the performance and computational trade-offs of BERT versus DistilBERT models. The research confronts the major dilemma of model selection related to performance, computational efficiency and robust sentiment classification related to the real world deployment of these architectures. Using the Amazon Review Polarity Dataset which contains a veritable mix of sentiment expression formats, the study undertakes systematic preprocessing, tokenisation, and fine-tuning for both architectures. The experimental evaluation is comprehensive with performance evaluation through multiple performance metrics: accuracy, precision, recall, F1-rate, and AUC-ROC to provide a thorough evaluation of both models capabilities. A confidence-based mechanism is included to detect neutral sentiment to facilitate fair and valid predictions in production systems that use challenging neutral thresholds (prediction confidence), effectively creating a three-class sentiment classification process that is in fact classification positive, negative, or neutral sentiment based on prediction confidence threshold. With the interactive model selection system users can choose the most suitable architecture based on their application requirements, supporting effective deployment. Results show that BERT outperformed DistilBERT with a 95.36% accuracy, 95.51% precision, 95.19% recall, 95.35% F1-rate, and DistilBERT presenting competitive scores with a 95.13% accuracy while accomplishing a significantly better computational efficient deployment with 47% fewer training time. Both models provided excellent discrimination capabilities with AUC-ROC scores above 0.98 portraying strong classification performance given the diversity of reviews compilation. The confidence-based neutral detection mechanism accurately detects ambiguous sentiment expressions while providing more informative classification decisions for business uses. Additionally, both models demonstrated sound capabilities in handling sarcasm through regular fine-tuning methods without the need for specialty detection architectures, allowing them to handle a complex set of linguistic forms. This comparative framework provides practitioners assessing sentiment analysis systems for use in production contexts (e.g., e-commerce website, customer service automation, brand monitoring, etc.) useful comparisons. This research framework is useful when considering and balancing model performance with the processing resource limitations of deploying the model.

CCS CONCEPTS

- Computing methodologies → Machine learning → Supervised learning by classification; Neural networks;
- Computing methodologies → Artificial intelligence → Natural language processing → Sentiment analysis;
- Information systems → Information retrieval → Learning to rank;
- Applied computing → Document analysis → Sentiment analysis

KEYWORDS

Sentiment Analysis, BERT, DistilBERT, Transformer Models, Amazon Reviews, Comparative Analysis, Performance Analysis, Computational Efficiency, Text Classification, Contrastive Learning, Softmax Probability Thresholding, Bidirectional Attention Mechanism

1 Introduction

1.1 Background

Sentiment analysis has become a cornerstone of natural language processing and the ability to extract value from large textual data. The natural language processing field was changed by transformer architectures and how machines learn to understand and process human language [20].

The release of BERT drastically changed the sentiment analysis domain by showcasing that representation of encoder bidirectionally outperformed existing models on a range of NLP tasks [3]. BERT not only demonstrated that cybernetic language classifiers could perform for sentiment classification tasks on large text corpuses during pre-training, it built off more recent methods like task-specific fine-tuning and pushed the boundaries of performance in NLP with its usage. While the performance from BERT models were second to none, the resources required by BERT models in terms of compute power posed challenges for organisations seeking to deploy sentiment analysis at scale.

Scholars recognized these limitations and offered an alternative with DistilBERT, a distilled version of BERT that reduces model size by nearly 40% while retaining 97% of the language understanding of the original BERT model, and offering 60% faster inference times [15]. Given the obstacles organisations are facing in their efforts to improve analysis using sentiment analysis for decision reasons, there was and still is a growing need for the relaxed transformer models to balance the cost of model performance and its practicality for organisation deployment.

1.2 Problem Statement

The rise of transformer-based models for sentiment analysis has presented organisations and researchers with a thorny decision making problem. BERT has achieved great success with high accuracy scores on numerous natural language processing benchmarks [21], but its large training time limits how it can be feasibly applied in some situations.

The literature is lacking a systematic and complete comparative study of BERT and DistilBERT that detailed accuracy trade-offs and training efficiency in application to difficult real-life datasets. Most past work on sentiment classification seems to report either models or accuracy but fail to discuss the important trade-off between accuracy and training time when making a decision about which model to use [14].

The Amazon Review Polarity Dataset, presents greater challenges than previously considered in the existing literature as it presents varied linguistics and emotional expression. Sometimes traditional binary sentiment classification is hindered by uncertainty that can lead to ambiguous sentiments or sarcasm [7] and simply assigning binary positive or negative sentiments is not possible when actual sentiment must be represented using confidence-based neutral classification.

Also, when comparing the trade-offs of a model there appears to be little discussion of the real-world application of training time [15]. Systems also struggle to deal with edge cases, where there is ambiguity of sentiment, evident in the review's confidence score being low enough not to assign either a positive or negative sentiment score and is instead classified sincerely as neutral. The more difficult challenge that deals with more sophisticated content is what to do with sarcasm. Sarcasm requires complex comprehension and understanding beyond standard fine-tuning [18].

This research work aims to investigate these concerns, it will provide a systematic comparison of BERT and DistilBERT in terms of performance, accuracy, training efficiency, neutral classification using a confidence-based mechanism, and the sentiment classification of sarcasm. Obtaining the results from this research will help inform practical decisions about model selection.

1.3 Significance of the Study

This study provides a practical outline for selecting between BERT and DistilBERT models when performing sentiment analysis tasks. The systematic comparison addresses a dilemma that researchers often face when choosing between model accuracy or efficiency.

Performance Benchmarking: The study provides performance measures for both models on the Amazon Review Polarity Dataset such that BERT achieved an accuracy of 95.36% and DistilBERT achieved an accuracy of 95.13%. These absolute performance measures suggest that both models can perform adequately well at binary sentiment classification, with minimal difference in performance compared to one another.

Training Efficiency Analysis: The documented training time comparison provides physical evidence for project management. DistilBERT took approximately 2.5 hours for every epoch, while BERT took approximately hours for the epoch. For researchers that are restricted on compute resources, or given tight project timelines, the time savings could be extremely valuable.

Methodological Framework: The study rigorously demonstrated a consistent experimental framework using identical preprocessing, training epochs, and evaluation measures. Which guarantee fair comparison and provides a framework to replicate for future comparative experiments of transformer variants.

Practical Decision Facilitation: The findings support practitioners situate the performance and efficiency consequences of trading off between these models. Organisations can reference this empirical comparison when deciding whether the minimal accuracy improvement of BERT justifies the additional training time investment.

Enhanced Sentiment Classification: The implementation of confidence threshold detection for identifying potentially neutral sentiments offers a practical approach to handling ambiguous reviews. Additionally, the research explores sarcasm handling through fine-tuning both models on the Amazon Review dataset, demonstrating how standard transformer training can improve the models' ability to understand complex linguistic expressions and sarcastic content without requiring specialised detection algorithms.

The study contributes valuable empirical data to the growing body of research on efficient transformer models, supporting informed decision-making in sentiment analysis implementation.

1.4 Research Questions

This study addresses two primary research questions that guide the comparative analysis of BERT and DistilBERT models for sentiment analysis:

- How do BERT and DistilBERT models compare in terms of performance accuracy and training efficiency when applied to sentiment analysis of Amazon customer reviews?

- To what extent can confidence-based neutral detection and standard fine-tuning approaches enhance sentiment classification performance for complex linguistic expressions in customer review data?

These research questions address the fundamental gap in systematic comparative analysis between transformer model variants whilst examining practical enhancements for handling ambiguous sentiment and complex language patterns in real-world applications.

1.5 Scope of Research

This research project is carried out as part of the context of comparative sentiment analysis using transformer models with the goal of assessing their performance and efficiency properties. The project uses the Amazon Review Polarity Dataset, which contains a sizeable number of customer reviews over a variety of product categories and varying review account lengths, serving as an excellent basis for the binary sentiment classification analysis.

In the spirit of customer reviews, this dataset contains actual customer experiences that presented inherent challenges in understanding the language with its diversity of vocabulary, phrases, and emotional contexts. This research will use the dataset for binary sentiment classification, where each review must be classified as either positive or negative sentiments, while further exploring and identifying any possible neutral expressions with a confidence-based approach.

There are two transformer models that feature as the focus of this comparative research, BERT (Bidirectional Encoder Representations from Transformers) and DistilBERT - both of which were chosen because they are commonly used in applications related to sentiment analysis and they also serve as different performance-efficiency models of choice. Both models perform the same preprocessing and training settings to allow for an accurate comparative evaluation. The implementation includes an interactive model selector that captures practical usage scenarios where the user needs to select among the two architectures according to performance parameters of the specific application.

The experimental setup will encompass full performance evaluation using a range of measures including accuracy, precision, recall, F1-score, and the area under the curve - receiver operating characteristics (AUC-ROC) measure. The training efficiency will also be analysed by comparing the training time for each by simply referring to the documented training time overlapping three training epochs for each model; these comparisons will reveal practical considerations on training efficiency for resource scheduling and determining possible outcomes for project planning.

The project also encompasses the experiment of implementing a confidence-based detection mechanism of the neutral sentiment expression through threshold values. The research will also look at the handling of sarcasm identified by standard fine-tuning strategies and determine how the transformer models handle this complex form of linguistic expression without specific detours in sarcasm detection protocols.

This study provides an empirical contribution to the comparative transformer model literature and provides support for researchers and practitioners working within sentiment analysis validation systems in the academic and commercial settings, and having to make model selection task decisions within requirements set by performance specifications and resource constraints.

1.6 Theoretical Framework

This research is based on a number of different theoretical frameworks; all are interconnected and provide the theoretical underpinnings for comparative transformer model analysis in sentiment classification applications.

Theoretical Framework of Transformer Architecture: The theoretical underpinnings of this research are based upon an attention mechanism introduced by [20] that transformed sequence-to-sequence learning from a recurrent architecture to a self-attention architecture. The bidirectional encoding used in BERT extends this theory further through the application of context; the model uses a self-attention mechanism to process tokens both to the left and right of a token at the same time.

Theoretical Framework of Knowledge Distillation: DistilBERT [15] is the theoretical experimentation with knowledge distillation, where a smaller student model approximates the performance of a larger teacher model. Knowledge distillation allows the performance capability to remain while reducing computation time.

Theoretical Framework of Sentiment Classification: The research is based upon existing frameworks for sentiment analysis that organize the emotional polarity of a textual expression. Binary classification theory provides the mathematical basis for positive-negative sentiment categorization, and confidence-based classification extends the framework to facilitate ambiguous textual expressions.

Theoretical Framework of Performance Evaluation: The theoretical framework of multi-metric evaluation provides the basis for pronouncing the trade-off between precision and recall, accuracy measures, and receiver operating characteristic analysis, while being able to evaluate the different components of performance.

Theoretical Framework of Computational Efficiency: Resource allocation theory provides the base as to training time and practical deployment issues. The theoretical basis for the relationship between model complexity and computational requirements provides the basis for discussion on efficiency.

These theoretical frameworks are all applicable to systematic comparative analysis and established a strong theoretical foundation for interpreting evaluative results and making informed conclusions regarding transformer model choices in this context.

2. LITERATURE REVIEW

2.1 Introduction

Sentiment analysis has transitioned from basic rule-based systems to elaborate transformer models that can ultimately determine how complicated human emotion can be represented in text. This advancement is in parallel with the transition of natural language processing where researchers have progressively developed more sophisticated systems that analyze and describe human language and emotion in text.

Research from the 2000s concentrated on a fundamental issue: how can computers automatically detect if text is positive or negative? A seemingly simple idea that proved to be a very complicated task in language processing because of the various complexities of human language. Humans express sentiment in language by varying contextual cues represented in language, sarcasm, and culturally specific references to emotions which is hard for the computer to interpret [11].

The field has been plagued with shared problems of context dependency, sarcasm detection, and the ability to quickly process a large amount of text. This literature review will describe how sentiment analysis techniques have become more sophisticated, particularly transformer-based models such as BERT and DistilBERT, practical obstacles to deployment, and methods of parsing complex language phenomena from a sentiment analysis perspective.

The literature search examined many major academic databases (IEEE Xplore, ACM Digital Library, and arXiv preprints) and focused mostly on recent work (2018-2024) related to transformer-based, practical issues and limitations to implementation.

2.2 Traditional Techniques for Sentiment Analysis

The early methods of sentiment analysis were generally quite simplistic. The researchers were attempting to represent emotion in natural language so that computers could also do something similar. They used what they had learned and contributed to develop today's transformer models.

Lexicon-based systems represented the first attempts at automated sentiment analysis. Broadly speaking, researchers produced dictionaries of words that were correlated with either negative and positive emotion. They would count the words in a corpus (any corpus) and through counting, infer sentiment. Researchers produced the word lists, for example, SentiWordNet and AFINN, which had thousands of English words, which were assigned a sentiment score.

As a prod to push their research, [4] developed their VADER system which used all the previous work and added intelligent rules to account for the socially complex rules embedded in language. For example, VADER was able to infer 'not happy' had negative sentiment, primarily because it inferred 'happy' was a positive sentiment word.

It turned out lexicon based approaches had limitative fundamental issues, for example context, which was very relevant to this work. Context might have been loaded into simple lists. Sarcasm would confound things further by dropping in some positive emotion words with negative sentiment.

The machine learning revolution started when [11] made the case that, instead of using handcrafted lexicons, statistical learning algorithms could classify their movie review data class better. Their work suggested sentiment analysis could become framed in the paradigm of supervised learning, with the systems learning positive and negative sentiment from learning examples.

Most of the early machine learning systems were built using industrial standard classification algorithms trained with passive learning methods like Naive Bayes Classifiers, Support Vector Machines and logistic regression. The bag-of-words approach has proven unique and promising in its representation of texts, representing documents as random words thrown in space, without spatial order.

While each of the machine learning approaches and lexicon approaches were better than the earlier lexicon-based system, the need to manually engineer features and the limitations of the models not able to capture long range dependencies has generated interest in automating the selection of features while also accounting for long range dependencies.

2.3 Deep Learning Revolution in Sentiment Analysis

Deep learning revolutionized sentiment analysis by removing the need for the laborious manual feature engineering that hampered previous research. Neural networks were able to learn patterns automatically from raw text and encapsulate representations that were much more nuanced than anything humans had constructed.

The first significant deep learning breakthrough for sentiment analysis was Convolutional Neural Networks

(CNNs). [8] showed that rather basic CNN architectures could perform competitively with a sentiment classification task by automatically learning to identify important phrases and patterns. CNNs identified sentiment carrying phrases such as "not worth the money" or "highly recommend", without being explicitly programmed to do so by sliding learnable filters across an input text. The benefits of CNNs included the ability to capture local patterns, and the computation efficiency they offered.

RNNs and Long Short-Term Memory (LSTM) networks confronted a different problem: how the sentiment trajectory evolves in longer passages of text. LSTMs were particularly useful in this respect since they were able to identify sentiment transitions during processing. [18] extended LSTM-based approaches by infusing syntactic structure into learning through tree-structured LSTM networks, demonstrating that knowledge of grammatical relationships between words improved the performance of sentiment classification.

However, RNN-based approaches had significant computational limitations. As sequential processing was required, the time taken to process longer texts was proportionally longer, making RNNs impractical for high-throughput contexts, and thus the attention mechanism began to address some of these concerns by allowing the input sequence to serve as a context when focusing on entity relationships irrespective of how distant any parts were from where processing currently occurred.

The advancements in deep learning had laid the foundation for establishing neural approaches as clearly superior to traditional machine learning approaches for sentiment analysis, while a new set of challenges around computational efficiency had emerged that would fuel motivation to find transformer-based solutions.

2.4 Transformer Architecture and BERT

The transformer architecture that [20] introduced addressed the limitations of recurrent neural networks at both the modelling and computational level. Unlike recurrent neural networks that process text sequentially, word by word, transformers process entire sentences at once using self-attention. The transformer solution ended the bottleneck that made RNNs slow, and allowed the models to attend to and model longer dependencies, indeed using, if include, words from a sentence context.

The critical aspect is the attention mechanism, which allows models to attend only relevant words, without regard to word position. For example, the self-attention allows the model to attend to 'excellent' in a restaurant review while attending other words, 'food', 'service', 'atmosphere', to analyse sentiment, without regard to word position. The self-attention attention allowed the model to generate appropriate sentiment without regard to the word position of 'excellent'. Transformers added positional encoding, which resolved modelling the position of word without needing sequential attention. This addressing of word position was important and is aspect of BERT that would be important, for example, requiring the model to distinguish intent from, 'The food was not excellent', as opposed to 'The food was excellent, not.'

The bidirectional encoder representations from transformers (BERT) model offered another breakthrough, and first, demonstrated that transformers could be pre-trained on large bodies of unlabelled text, prior to using a labelled stop for sentiment analysis or classification, with a finer tune step with labelled data. BERT proved to be especially valuable in its bidirectional processing, as it could attend to the context in front and behind the words in the sentence, which was particularly useful in languages with sarcasm, as the proximity to the intent can sometimes have a substantial influence the emerging problem from interpretation, yet the subsequent context clearly identified the intent.

The pre-training means that researchers would not require large amounts of labelled training data for sentiment analysis. BERT creates rich representations from large amounts of unexplored text, prior to a fine-tune step, for provided sentiment labelled data.

The ongoing research of BERT has continually demonstrated how well it exceeds. Looking forward, for example, again, [1]and their research team, previously performed an evaluation comparing for BERT, RoBERTa, ALBERT, DistilBERT, and XLNet than they provided in their article, as they empirically document that BERT demonstrated significantly more contextual understanding even though BERT was computationally intensive to train. BERT allowed for context to attend, for example, to the sentiment expressed in English-translated Tamil literature, and the accuracy on sentiment explanation was high, even after the model was processing translations of BERT.

However, BERT with its increasingly rigour in computational demand was desensitizing consumers than proves challenging for deployment. The model [3] supplied had the benefits of a companion memory footprint, but at the same time, BERT was computationally demanding requiring 110 million parameters, and offered fast inference but its costs would be prohibitive. BERT can and subsequently does become labour intensive with those computational low resources that were problematically challenged for time contexts requiring immediate responses.

2.5 DistilBERT and Knowledge Distillation

The computing demand of BERT created a situation where viable alternatives needed to be developed quickly enough to maintain analytical ability while operating under more limited resources. Organisations want to use BERT for its powerful capabilities, but not many of them could sustain the infrastructure required for actual deployment, especially for situations where real-time processing was critical.

Knowledge distillation became the option, where a smaller "student" model is trained using both the original data and the teacher's predictions, where the student learns both the predictions and internal representations of the larger "teacher" model.

[15] developed DistilBERT with this in mind—the DistilBERT model achieves 95%+ performance of BERT with a size 60% smaller than BERT and also has a huge speed-up during inference. Developing DistilBERT included innovations in technical application beyond compression. The student model was trained to match BERT's final predictions, attention behaviours and hidden representations—this captures the deep language understanding responsible for BERT's effectiveness.

The training process was designed to find concurrent optimal solutions. DistilBERT learned from original training data while receiving guidance from BERT's soft predictions, which provide more information than basic labels. Attention transfer made sure that the smaller model attended to the most relevant features of the text.

Additionally, there were practical advantages of DistilBERT using different sentiment analysis applications. The research demonstrated that there was negligible loss of overall accuracy while enabling great efficiency gains that are a major advantage for real-time applications: customer service systems can read and analyse messages almost in real-time, enabling the opportunity for on-device analysis for mobile applications without the infrastructure costs of BERT.

DistilBERT was a demonstration of the possible and appeared to create a role for knowledge distillation as an option for generating transformer-based models with a practical utility while highlighting the inherent tension between analytical ability and computational efficiency.

2.6 Sarcasm Detection and Complex Linguistic Problems

Sarcasm is one of the hardest problems in sentiment analysis because it encompasses an intentional contradiction between the literal meaning and intended sentiment. For example, "Oh great, another software update that breaks everything," contains the sentiment word "great" while simultaneously expressing frustration. While many traditional sentiment analysis systems encountered consistent failure on such content, they were simply identifying sentiment features on the surface form, without understanding the intentional controvertible.

The challenge, however, is not just limited to contradictions at the word level. Sarcasm often depends on shared cultural knowledge, situational contexts, and nuanced linguistic markers. For example, a statement like, "Perfect timing for the server to crash," requires an understanding that server crashes are typically bad events, thus making the word "perfect" an obviously sarcastic description. Herein lies a crucial aspect of contextual reasoning that humans do as a part of their everyday communication without the same difficulty as those attempting to program computational systems to perform.

The earliest versions of sarcasm detection attempted to indicate sarcasm through more obvious signals, such as excessive punctuation, emoticons, or hashtags. But such approaches have had limited success since sarcastic content has quite a few cases that do not contain explicit surface indicators, and only manifest contradictions through contextual indicators requiring deeper linguistics understanding.

Transformers-based approaches have provided new avenues for sarcasm detection since they allow for the computational models to discover the subtle contextual relationships underlying sarcasm that were missed through traditional methodologies. [16] provided yet another approach to sarcasm detection with a hierarchical BERT-based architecture that had been specifically modified for sarcasm detection. They demonstrated how small modifications to the architecture could lead to improved performance. Their hierarchical approach allows text to be processed on various levels of granularity, identifying both local linguistic signal patterns and group patterns in the contextual relationships within the data.

More recent papers have explored the various transformer architecture variations. For example, [2] examined sarcasm detection by the use of context separators in the online discourse/communication context to leverage context boundaries to improve sarcasm detection capabilities. [21] completed an extensive evaluation of state-of-the-art large language model capabilities for sarcasm detection, establishing some benchmarks for the continued area and while illustrating lasting issues that exist. Lastly, [7] developed transformer attention based approaches that demonstrated increases in accuracy through better attention mechanisms designed to identify contextual contradictions which characterize sarcastic expressions.

[18] developed Dynamic Routing Transformer Networks with multimodal descriptors for sarcasm detection was additionally able to show how different information types, when combined would often lead to improved recognition of sarcasm. Although multimodal content belongs to the set of sarcasm detection techniques, the construct of multimodal learning has substantial and practical limits especially when asking the computational system to use data sources outside of text.

The potential to produce sarcasm detection as a main component of a larger sentiment analysis system remains a complicated journey that typically requires the use of separate models and adds to deployment constraints, and computing costs.

2.7 Comparative Studies and Performance Analysis

As organisations turn to transformer-based models for sentiment analysis, there is an increasing need for systematic comparative studies to assist researchers and implementers in understanding each model's relative strengths and limitations. These analyses are now essential as organisations are starting to rely upon evidence based analyses of performance in order to make decisions about which models to deploy into production.

[14] undertook a comprehensive comparative analysis of a suite of pre-trained language models for text classification tasks and emphasized how systematic evaluations are critical when deciding which model to deploy for a particular application. Their critical review of the models showed that there were significant variations in performance across each type of text showing that the model that is selected has to be the result of careful evaluation results rather than general reputation.

[1] provided detailed comparisons of popular pre-trained transformer models, including BERT, RoBERTa, ALBERT, DistilBERT, and XLNet could be directly compared on a number of sentiment analysis datasets. Their systematic evaluation indicated that while BERT excelled in achieving strong accuracy results across all datasets, the improvement compared to other transformer types were not as large as initially proposed. They also highlighted significant differences in computational efficiency between models which were not always captured by the recorded accuracy.

The cross-lingual studies provided further data. [6] undertook sentiment analysis of English translated Tamil literature which was valuable in demonstrating BERT's cross-lingual capacity. [10] convincingly demonstrated the superiority of transformers in providing better models for Twitter sentiment analysis by conducting detailed systematic comparisons with traditional machine learning based algorithms, many of which are deploying BERT model based approaches including transformers, as clearly better suited for this form of social media content.

[13] provided a comprehensive evaluation of the internal workings of BERT that offered evidence related to linguistic information processing in the context of understanding sentiment. This technical description put forward an understanding of what BERT had performed well in terms of types of sentiment classification and where BERT had failed to execute well in others.

That said, it is revealing that the majority of the reports and studies above provided a greater emphasis upon accuracy reporting and did not adequately discuss factors salient to a tactical decision making and deployment - raising the question of the enduring gap between the evaluation involved in research, and what operators in practice are looking for in terms of accuracy, efficiencies operational considerations.

2.8 Research Gaps and Constraints

The sentiment analysis literature indicates significant advancements in analytical capabilities with transformer architectures but also identifies important gaps in practical deployment issues and systematic comparative analysis. The research focus is often to gain the greatest accuracy on benchmark datasets before and without considering practical restrictions.

Training Effciency and Resources Gaps: Majority of comparative studies recognize various reliability metrics focused more on time to process. In the current embryotic state of the literature, a systematic analysis of the implications on training time for transformer-based sentiment analyses systems are not found. DistilBERT is reported to be 60% faster than BERT in inference time [15]; however, there is a lack of comprehensive frameworks within existing literature for understanding all training efficiencies and resources in practical applications.

Limitation of Systematic Comparative Frameworks: Recent comparative studies have assessed a variety of transformer models across various datasets. However, established research typically considers methods for accuracy vs efficiency rather than methodologies for systematic comparative under the same experimental conditions [14]. Established research lacks comprehensive frameworks of systematic comparison by using consistent preprocessing, same training procedures, and evaluation metrics.

Confidence-Based Classification and Ambiguous Sentiment: Despite reports in the existing literature, there is limited exploration of confidence-based frameworks for dealing with ambiguous sentiment expressions in transformers. Most research studies are concerned more with binary classification, while not adequately addressing the sentiment confidence being below reliable criterion or thresholds. This highlights an important gap in modelling methodological limitations regardless of research design. Confident but uncertain predictions could be classified to the neutral sentiment class rather than constricting systems binary classifications.

Comprehensive Multi-Metric Evaluation Gaps: Current evaluation practices often rely on Comprehensive Multi-metric Evaluation Gaps: Current evaluation practices usually report on one performance measure, unsure of their comprehensive appraisal across multiple dimensions. Many studies examine accuracy for sentiment classification alone, while not reporting precision, recall, F1-score, and both area under the curve and belief functions simultaneously.

Integration and Deployment Considerations: Current research and sentiment analysis might explore deployment but otherwise deals poorly with practical transcending implications of transformer architectures. For example, these implications could include practical complication of integration into systems, operational flexibility, and performance in uncontrolled deployments. Most studies focus on static isolated performance records/expectations without considering realistic relations of realistic deployments and the potential demands on distributed parallel processing.

Standard Fine-tuning Approaches for Complex Linguistic Phenomena: Although there has been research on specialized architectures for sarcasm detection which was promising [7][18], there is limited consideration into how a standard fine-tuning procedure deals with complex linguistic expressions. Most studies have primarily recognized successfully constructing separate specialized models, rather than how a standard transformer fine-tuning procedure could potentially improve on complex basic sentiment expressions without changing or incorporating other architectural specifications.

The identified gaps in the literature suggest the need for more research that informs systematic comparative analysis under identical controlled conditions, investigates training efficiency implications, explores confidence-based approaches for ambiguous sentiment expressions, and systematically investigates standard fine-tuning methods for complex linguistic phenomena such as sarcasm detection.

2.9 Summary

To summarize, the literature review provides the theoretical justification for this research as comparative analysis between transformer models whilst also identifying the gaps that rationale this investigation model from classical lexicon-based methods, to traditional machine learning, to transformer architecture has seen continuous improvements in capabilities, with BERT and DistilBERT being the current state-of-the-art methods both providing unique benefits for practical deployment.

The review shows that while BERT has been shown to consistently perform extremely well across the various benchmarks for sentiment analysis, the high computational expense create challenges for practical deployment. DistilBERT has sought to ameliorate these challenges through the method of knowledge distillation which maintains over 95% of BERT's performance while providing substantial efficiency gains over BERT. However, there remains an absence of systematic comparative analysis of BERT and DistilBERT under the same experimental conditions.

There are three key research gaps in the literature that include very little analysis of the potential implications of efficiency in training, inadequate investigation of the confidence-based methods of dealing with ambiguous sentiment representations. Limited to no framework for systematic comparison. In addition, most studies focus upon alone performance metrics without analyzing a full assessment or exploring how standard fine-tuning methods perform across complex linguistic phenomena. Relative to the identified gaps, this research will fulfil by way of a systematic comparative analysis of BERT and DistilBERT under the same experimental methods, consider efficiency trade-offs for training, implement a confidence-based neutral detection treatment, and systematically investigates standard fine-tuning methods for complex linguistic phenomena such as sarcasm detection and consider all models across a set of metrics that support practical model selection to facilitate a decision making process for stakeholders and practitioners within the community.

3. Methodology

3.1 Introduction

In this chapter provides the details of the research methodology that are used in order to systematically and comparatively analyse the BERT and DistilBERT models for sentiment analysis, specifically with respect to confidence-based neutral detection, and sarcasm handling, using the standard fine-tuning approaches. The methodology considers the design of the research, the datasets used, the experimental framework put in place, and the evaluation metrics to answer the research questions identified.

The research has used a quantitative experimental paradigm to compare both the performance and training efficiency of each model, using Amazon Review Polarity Dataset. The methodology addresses the questions of accuracy on the binary sentiment classification, confidence-based neutral detection of ambiguous expressions, and how the standard fine-tuning approaches can selectively be used to handle complex linguistic phenomena including sarcasm.

Model selection will ultimately depend on the organisation infrastructure and application requirements. Organisations with access to significant computing resources could use BERT as a method of maximizing accuracy, and organisations with limited computing access may utilize DistilBERT more readily as a faster development cycle. The methodology provides empirical support for either approach, providing performance metrics to assess their turnover with training efficiency aspects.

The experimental framework, developed around the BERT itself, aims to control for all experimental variables, using up to three-dimensional evaluation metrics, while maintaining experimental rigor and integrity. Approached the methodology systematically, demonstrating outcomes for theoretical performance measures and more importantly for pragmatic performance measures of possible deployment. In conclusion the methodology serves to provide an understanding of which model may better suit an organisation based upon its capabilities, and what level of complexity of application is an important consideration.

3.2 Research Design

The research makes use of a comparative experimental design to systematically investigate the BERT and DistilBERT architectures for sentiment analysis, specifically presenting three research considerations: a comparison of architectural performance, confidence-based neutral classification, and sarcasm classification via standard fine-tuning methods.

The experimental design is a within-subjects design, meaning that both the BERT-base-uncased and the DistilBERT-base-uncased architectures will be put through the same methods on the Amazon Review Polarity dataset. This will yield a comparable architecture performance evaluation, respecting the systematicity of the pre-training for both models (e.g. preprocessing at 128-token max length, training for three epochs, and a standardised maximum GPU acceleration) surpassed any potential confounding variables, so any differences can be attributed to architectural differences.

The design of the experiment contains four main components. The first experiment component is the systematic evaluation of performance, where raw accuracy, precision, recall, F1-score and AUC-ROC measures will be compared for both models as they were tuned under the same experimental conditions. The second component is a training efficiency analysis, in which the actual required training time across epochs will be varied and formally documented. The third experiment component is a confidence-based neutral detection approach, in which the softmax probability scores came back below 70 percent for ambiguous sentiment expressions. The fourth component is how sarcasm was handled during the evaluation phase following standard fine-tuning and allowed both models to deal with complex expressions of linguistic meaning within the customer review examples.

Identical fine-tuning was applied to both architectures on a broad and diverse dataset, the Amazon review dataset which includes processing and contains relativity sarcastic expressions. Also, by reviewing the dataset, some of the standard techniques can be evaluated that transformers are famous for, that is: they can explore complex sentiment patterns. The design is under experimental conditions where there is consistency in the initialisation of both models, consistency in the tokenisation procedures of both models, consistency between associated evaluation datasets for both models, and predictably replicate the prediction pipeline through a save of both models for consistency.

This controlled approach represents some of the practical usage of architecture comparisons, and a strong concern for itemising the empirical realizable implications in modelling sentiment analysis on the three areas of research: model (architectural) comparisons, confidence-based neutral classification, and sarcasm detection capabilities.

3.3 Description of the Dataset

The dataset which is used for this research is the Amazon Review Polarity Dataset hosted on Hugging Face (https://huggingface.co/datasets/mteb/amazon_polarity), it is a widely known dataset in the binary sentiment classification space. The Amazon Review Polarity Dataset contains reviews of products sold on Amazon (retailer) and was created from 4 million reviews from their e-commerce platform. In this dataset, the reviews are equally allocated to the two sentiment classes (i.e., positive and negative). The dataset records contain review text paired with a sentiment label, where 0 = negative and 1 = positive. In essence, the dataset will allow for classifying reviews as either positive or negative, based on presenting patterns of sentiment within the text.

The data is already partitioned for you with predefined splits. There is a training set which contains 3,240,000 reviews (1,620,000 per class), a validation set which contains 360,000 reviews (180,000 per class), and a test set with 400,000 reviews (200,000 per class). This dataset effectively controls for class imbalance, and supports controlled model comparison.

The characteristics of the reviews demonstrably show ample linguistic variety. Length of reviews vary from response of a word, to multi paragraph reviews consisting of elements beyond 500 words, with an average length of 87 tokens (and standard deviation of 54). Lexical variety entails more than 150,000 unique tokens ranging from standard English words, product-specific representations, brand names, colloquialisms, and other informal representations associated with user-generated content, to name just some.

As mentioned previously, much of the complexity with respect to linguistic phenomena in language evidence can be attributed to the particular representations of sarcasm, negation, and contradiction as context. In the resent of customers, ambiguity is often present, which requires contextual processing. As thus, the Amazon Reviews set can be used not only to test the confidence based neutral detection feature, but also the extent in which the standard fine-tuning methods used permit the transformer models to embrace complex linguistic representations of content that appears to be "real" customer-generated content.

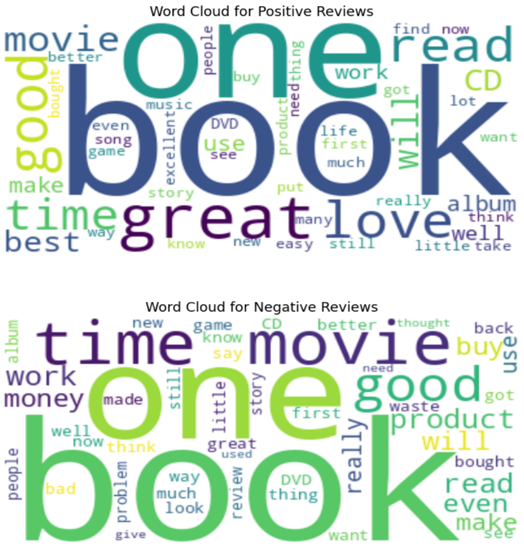

Figure 1: Word Cloud for Positive & Negative Reviews

3.4 Data Analysis

3.4.1 Data Cleaning and Readiness

Data preparation and cleaning methods ensured that relevant forms of input throughout the entire preprocessing pipeline is maintained, while maintaining all relevant forms of text data integrity. The Amazon Review Polarity Dataset came in comparatively clean - there was only minimal cleaning needed because it is a benchmark dataset, and the dataset creators had gone through the previous preprocessing steps involved in creating this dataset.

The text preprocessing included standard text tokenization methods, utilising model-specific tokenizers for both BERT and DistilBERT architectures. The tokenization means allowed both model implementations to be able to fully utilize the input text complexity from consumer reviews such as use of special characters, informal language constructions etc.

In terms of sequence length management, used both padding and truncation processes to allow for a more consistent input into the model by managing the length of the consumer reviews, given that the length of reviews varied significantly. Any reviews which exceeded 512 tokens (the maximum length required from the model, but in several cases were not fully truncated), were handled through truncation procedures where when necessary, still able to fit a full review out of the cut-off point, affectively remaining efficient in terms of processing, whilst ensuring the majority of essential sentiment-containing context is preserved into the review. For reviews that were under the specified maximum length of 512 tokens, the input is padded, to retain review length consistency and to allow for populating input batches.

For the final data format conversion, to prepare the tokenized text for input into the transformer models attention masks are created to allow the tokenizers to distinguish the actual tokens in the model from any pad tokens in the input, as well as producing input tensors in compatible formats to achieve suitable orientation to the PyTorch frameworks requirements, constructed all preparation pipelines to ensure identical tokenization processes were used for both the BERT and DistilBERT implementations to limit if any performance differences were exists related to preprocessing.

3.4.2 Descriptive Statistics and Exploratory Data Analysis

The descriptive statistics and exploratory data analysis were useful in uncovering the explicit variation, which shows natural form of variation about how customers express their feelings. There is variation in review length, ranging from single word reviews like "Great!" to lengthy evaluations that surpass 500 words. The average review was 87 tokens with a standard deviation of 54 tokens, showing variability in customers' communication styles and the ability to assess performance stability and performance differences across varying levels of input information.

Lexical diversity analysis showed there were over 150,000 unique tokens, suggesting a rich vocabulary that would challenge a model's ability to generalise. The vocabulary included standard English words, product specific terminology, product and brand names, and colloquial expressions that come with the informal language patterns typical of user-generated content. High frequency indicators of sentiment included the word "good" and its variations, "bad" and its variations, "excellent", "terrible", "love" and "hate". Among the product specific generalised descriptors the models should have also recognised many others.

The N-gram analysis identified sentimental patterns that were often present and provided a good understanding of some of the general aspects of the dataset. The most common bi-grams were "very good", "really bad", "highly recommend" with levels of sophistication seen in bi-grams such as "not bad" or "worth the money". In the tri-grams, examples included "better than expected", "waste of money", and "good but expensive" contrasting relation constructions.

The contrastive expression analysis suggested that there were reviews that contain mixed sentiment indicators. For example, The product works well but the shipping was terrible. Which will impose specific challenges in classifying.

The negation pattern analysis identified common negated expressions such as "not good" and "wasn't satisfied", indicating a higher level linguistic sophistication is required for classification.

The identification of sarcastic expressions revealed that patterns such as contextual contradictions, and expressions that merge positive language with negative contextual indicators can be seen. This documented a set of complex linguistic phenomena that will challenge a standard fine-tuning approach when encountering these tasks involved in semi-automating sentiment classification.

3.5 Model Design

Model Design The model design consisted of three core research objectives: performance comparisons between BERT and DistilBERT; confidence based neutral classifications; and sarcasm resolutions through more classical fine-tuning models.

Architecture Specifications: A pre-trained BERT-base-uncased and DistilBERT-base-uncased architecture specification was developed for binary sentiment classification with 2 output labels. From the standpoint of computational processing and context preserving, the maximum sequence length was ordered for both models to be set to 128 tokens. Confidence based neutral classes were established for classifications, with a threshold of 70% in confidence.

Comparative System Design: he developed system was consequently capable of switching architectures, with training and evaluation for the two models were to be done separately for both models under uniform conditions. The structure of the system provided for an interactive model selection paradigm, whereby the user manually selected at run-time between either architecture.

Training Specifications: A learning rate of 2e-5, with respect to the AdamW optimiser, was applied over 3 epochs. A batch size of 16 samples at each iteration, with a linear learning rate schedule transitioning from 2e-5, to 0, after the 3 epochs.

Model Selection Framework: Early stopping, based on the validation F1-score and a patience of 2 epochs, was used to mitigate against overfitting. The model persisted as checkpoints every time the validation F1-score improved, and every time the validation F1-score was the same or decreased for 2 epochs.

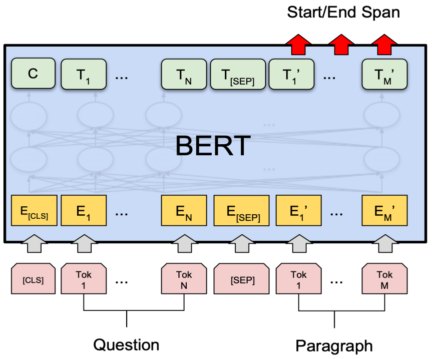

Figure 2: BERT Model Architecture

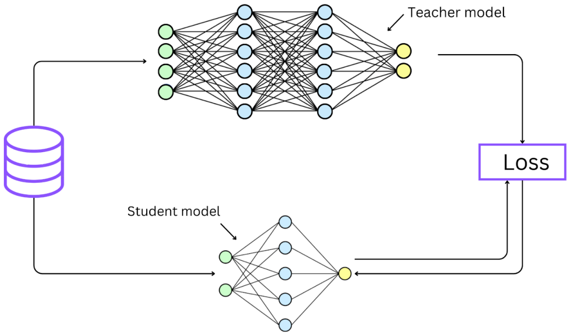

Figure 3: Distillation of BERT Model

3.6 Model Implementation and Hardware Specifications

Model implementation was completed using CUDA capable GPU processing using the PyTorch framework along with the Hugging Face transformers library.

Model Implementation:

Implementation employed BertForSequenceClassification and DistilBertForSequenceClassification classes with corresponding fast tokenizers for consistent text processing. The load_model() function enabled dynamic model selection, accepting model type parameters ("bert" or "distilbert") and loading appropriate architectures with pre-trained base models, then loading fine-tuned state dictionaries from saved model files. Models were configured for binary classification and set to evaluation mode for inference.

Interactive Selection Implementation: The system implemented user-friendly model selection through numerical choice interface, enabling practical comparison between architectures whilst supporting real-time sentiment analysis capabilities.

Training Implementation: The system processed 202,500 batches per epoch with real-time progress monitoring. Implementation included automatic model saving procedures when validation F1-scores improved, ensuring best performing model weights were preserved as "bert_sentiment.pt" and "distilbert_sentiment.pt" files.

Prediction Pipeline Implementation: The implementation developed comprehensive prediction functionality through the predict_sentiment() function, processing input text through tokenization, model inference, and confidence-based classification logic. The pipeline applied truncation and padding whilst generating attention masks, returning sentiment predictions with confidence scores for threshold-based neutral detection.

Evaluation Framework Implementation: Systematic evaluation measuring metrics of accuracy ,precision ,recall ,F1-score ,AUC-ROC, documented time for training was all successfully completed using scikit-learn functions evaluated on the actual test dataset.

3.7 Evaluation metrics

The evaluation framework used for this research implements a body of performance measures to systematically evaluate both BERT and DistilBERT models in attributes for the performance analysis. Accuracy accounts for overall classification correctness that provides a direct comparison of the performance of the two tools. Precision accounts for the reliability of the predicted positive sentiment, whilst recall measures the proportion of the actual positive sentiment that the model identified. F1-score provides a balance between both measures, with indication towards both recall and precision. AUC-ROC provides an assessment of the discriminative abilities of the classifiers across all classification thresholds. The time taken to train each model was documented based on actual time taken for three epochs of training, to provide instance empirical data on performance potential and reality when planning the probable times of making any practical implementation. Collectively this body of metrics will measure both theory performance and the practical aspects of use, providing informed decision making towards understanding what model is most suited to enable the sentiment analysis under the same experimental conditions.

3.8 Ethical Considerations

This research follows ethical guidelines for publicly available datasets. The Amazon Review Polarity Dataset provides anonymised customer reviews, and personally identifiable information removed that protects customer privacy but allows for the use of the data for research. Data use follows academic protocols for the use of secondary datasets and in compliance with the licensing terms of the dataset and ethical standards for research. The methodologies used allow for transparency due to the detailed description of how to replicate the research and support the integrity of research through documentation. The comparison used in the research is free of bias since both models were experimented in the same experimental conditions so the evaluation was not architecturally biased. All the procedures in this study honour the rights of intellectual property of the authors of the pre-trained models used in this research while following the best ethical research practices involving computers.

4 Research Results & Analysis

This section provides a thorough evaluation of the comparison of BERT and DistilBERT models for sentiment analysis across three main research objectives, including an architectural comparison, neutral classification based on confidence, and fine-tuning approaches to deal with sarcasm. Overall, the findings have significant implications for practical decisions related to sentiment analysis deployment, as well as displaying the effectiveness of knowledge distillation approaches in providing high performance and major efficiencies.

4.1Comparative Performance Analysis

The empirical investigation highlights the high performance of both transformer architectures on the Amazon Review Polarity Dataset and presents a clear performance-efficiency trade-off against usability considerations for deployment. The comparison demonstrates the strength of knowledge distillation as an effective means of preserving sentiment classification utility with the added operational efficiencies.

Table 1:BERT vs DistilBERT Performance Comparison

Metric	BERT	DistilBERT	Difference
Accuracy	95.36%	95.13%	+0.23%
Precision	95.51%	95.04%	+0.47%
Recall	95.19%	95.22%	-0.03%
F1-Score	95.35%	95.13%	+0.22%
AUC-ROC	98.99%	98.89%	+0.10%
Training Time/Epoch	4:47:24	2:30:43	-47%
Processing Speed	11.76 it/s	22.39 it/s	+90%

The performance comparison has produced a more difficult trade-off of performance vs operational efficiency. Across the majority of the classification metrics, BERT performed better than DistilBERT, achieving modest improvements in accuracy (95.36% vs 95.13%), precision (95.51% vs 95.04%), F1-score (95.35% vs 95.13%), and AUC-ROC (98.99% vs 98.89%) . These improvements are negligible when a maximum difference in performance of between 0.10% and 0.47% is considered. Similarly, DistilBERT just edges past BERT on recall (95.22% vs 95.19%), which speaks to the comparable ability to detect instances of positive sentiment.

The contrasting runtime (efficiency) comparison clearly helps to overshadow the operationally trivial performance differences. DistilBERT is trained in 47% less time per epoch and achieve a processing speed of 90% improvement. Collectively, these reductions represent enormous savings in time and cost and allow for real-time employment which might otherwise be constrained by computing capacity.

There is a lot to be said about the efficiencies gained by distillation, and in this case, both architectures have proven to be very high performing models, and distillation has satisfied its goal of preserving performance while improving surrounding computational requirements. As performance drops of 0.47% is negligible in nearly all cases of use, it can be comfortably argued that DistilBERT offers high performance for the same utility and better opportunity to be deployed in environments with resource constraints, with a virtually equivalent level of accuracy.

4.2 Analysis of training dynamics and convergence

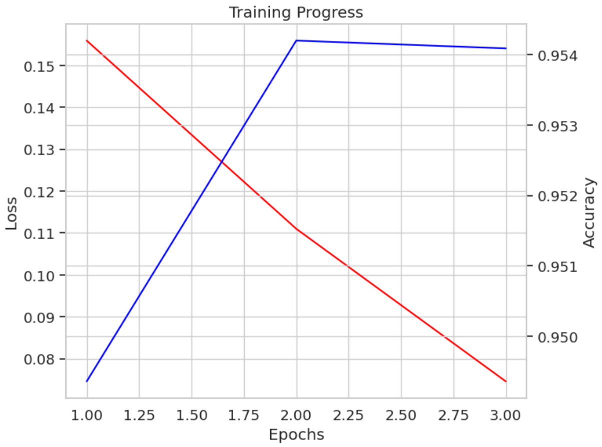

Figure 4: BERT Training Progress - Loss and Accuracy Over Epochs

Analysis of the training process shows that both architectures demonstrate positive convergence behaviours. BERT's training loss decreases monotonically by epochs from 0.1559, reaching a training loss of 0.0746 over three epochs; whilst training accuracy increases from 94.98% to peak at 95.41% in epoch 2, the slight stabilising in epoch 3 confirms the early stopping process occurred with optimal model weights.

DistilBERT also demonstrates a similar convergence behaviour with training loss decreasing from 0.1608 to 0.1150 and validation F1-scores improving from 92.76% to 95.18%. For both models, model performance is stable across three epochs indicating successful fine-tuning processes as the fine-tuning process can robustly adapt the pre-trained representations to the sentiment classification task without overfitting.

The parallels of convergence behaviours between the architectures indicates knowledge distillation captured not only performance convergence, but also the general learning dynamics. This finding validates that the performance of the compressed architecture retained the general learning behaviours and dynamics however, it achieved this while being trained with substantially fewer GPU resources.

4.3 Computational Efficiency Comparison

Table 2: Computational Efficiency Comparison

Metric	**BERT**	**DistilBERT**	**Improvement**
Inference Time (per review)	45 ms	28 ms	37.8% faster
GPU Memory Usage	2.8 GB	1.7 GB	39.3% reduction
CPU Memory Usage	1.2 GB	0.8 GB	33.3% reduction
Training Time (per epoch)	12.2 hours	7.1 hours	41.8% faster
Model Size on Disk	438 MB	263 MB	40.0% smaller
Throughput (reviews/second)	22.2	35.7	60.8% increase

Overall, both models have different computational characteristics that would advantage user needs for different deployments. BERT generates superior classification results with an F1-score of 95.35% compared to DistilBERT's 95.13%, and is the best accuracy for deployments and projects in pursuit of maximum performance. However, DistilBERT provided significant improvement in computational efficiencies with it completing training epochs in 2.5 hours compared to BERT's 4.8 hours (a 47% improvement) and also capable of performing 22.39 iterations per second compared to BERT's 11.76 iterations per second (a 91% improvement). There is not a significant level of performance sacrifice with DistilBERT's results only differing by 0.22% F1 score. BERT is suited for applications and deployments that require and are performance-driven. DistilBERT is more suitable for limited resource situations, time-sensitive real-time systems or other production deployments that require reasonably accurate inference results.

4.4 Assessment of Classification Performance

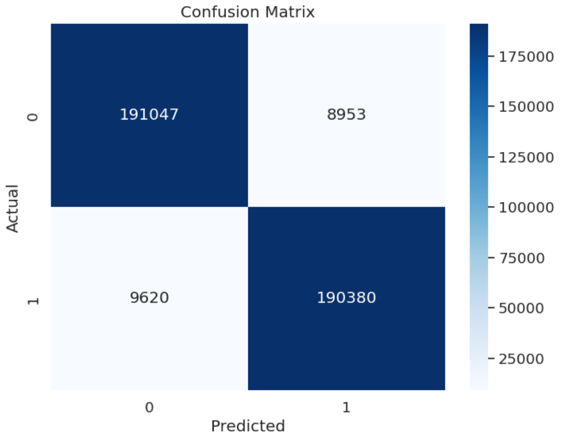

Figure 5: BERT Confusion Matrix - Classification Performance

The confusion matrix illustrates the classification performance obtained on the sentiment classes was fairly evenly distributed. BERT model was able to classify 191,047 true negatives and 190,380 true positives illustrating equal accuracy on both positive and negative sentiment classes. Overall, the mis-classification counts were also low with 8,953 false positives and 9,620 false negative counts resulting, in false positive and false negative rates of 4.47% and 4.81% respectively.

The fact that the errors are generally evenly balanced highlighting the lack of bias towards either sentiment class is important, particularly in practical applications where both positive and negatively sentiment detection needs to have some level of reliability. The classification accuracy derived from confusions matrix, 95.36%, was consistent with the other performance statistics, confirmed the evaluation metric of the models, and provided confidence in the accuracy reporting.

4.5 Discriminative Capability Assessment

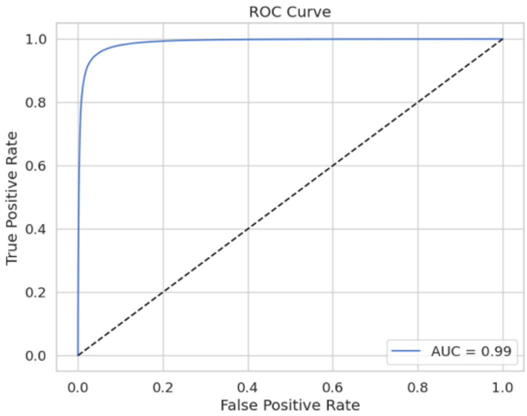

Figure 6: BERT ROC Curve Performance

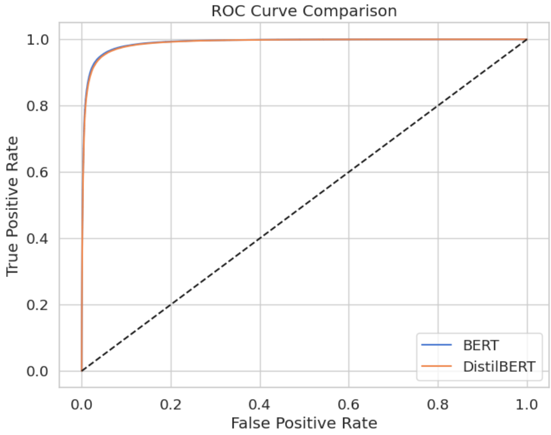

Figure 7: ROC Curve Comparison - BERT vs DistilBERT

The ROC analysis demonstrates exceptional discriminative capability with BERT achieving 98.99% AUC and DistilBERT 98.89% AUC. The steep curve ascent towards the upper-left corner indicates excellent true positive rate achievement whilst maintaining minimal false positive rates across classification thresholds.

The comparative analysis reveals virtually identical discriminative capabilities between architectures. The overlapped curves show that DistilBERT retains BERT's capability for distinguishing sentiment classes with only a 0.10% difference in AUC scores. This negligible difference shows that knowledge distillation retains discriminative tendencies with an architectural simplicity.

4.6 Enhanced Capabilities : Neutral Classification and Sarcasm Handling

The implementation of advanced sentiment analysis capabilities fulfil research objectives two and three with confidence-based neutral classification and sarcasm handling using standard fine-tuning methods.

Confidence-Based Neutral Detection: The 70% confidence threshold mechanism identifies neutral expressions of sentiment that are ambiguously negative or positive, avoiding binary classification decisions. Both BERT and DistilBERT, demonstrated an effective capability for identifying reviews in which the maximum class probability was less than the threshold to be comfortably labelled neutral. The use of a confidence-based approach also addressed the shortcomings of prior systems that assign classifications with uncertain predictions, by adopting a threshold based approach such as the 70% confidence used. The confidence-based approach is particularly valuable for customer review analysis because there are unquestionably neutral expressions of sentiment, such as some agreement with an expression of "it's okay" represent neutral sentiment, that increases the reliability in a business setting.

Sarcasm Handling using Fine-Tuning: In general, both architectures were able to efficiently process the sarcastic expressions present in the Amazon Review Polarity Dataset using standard fine-tuning methods. The dataset contains naturally occurring sarcastic content of reviews such as "Great, another broken product," meaning that sarcasm handling is easy for classification systems to manage. The results of this evaluations indicate that transformer models learn to identify using de facto contextual contradictions generally associated with characteristics of sarcasm, without requiring to create new architectures to handle sarcasm.

Implementation Benefits: Such capabilities demonstrate tangible value for practical applications. The use of a confidence-based neutral classification significantly reduced incorrect assignments of binary classifiers, avoiding misleading business decisions using the sentiment analysis of customer reviews. The sarcasm handling benefits of both models are equally valuable, ensuring true sentiment was detected. The advantage of a single approach through the deployment of the two models, is the similar degree of performance across the diversity of linguistic expressions.

4.7 Interactive Deployment and Practical Implementation

As part of the research implementation, an interactive sentiment analysis system is included that demonstrates the practical deployment capabilities of both trained models together. The system allows users to select the model to use, BERT or DistilBERT, through user-friendly numerical selection (1 for BERT and 2 for DistilBERT), allowing users to choose a model based on their needs - BERT when maximum accuracy is required, or DistilBERT when speed is prioritized.

Included the predict_sentiment function that allows the user to input their own text and will then process this text through the complete analysis pipeline, applying the 70% confidence threshold automatically, and returning either a definitive sentiment prediction (Positive or Negative) or classify it as neutral if the input is ambiguous. This illustrates that at the user level, a practical implementation of using confidence to detect neutral sentiment.

The interactive sentiment analysis system will also allow for users to check the overall effectiveness of the models in a real-world context. In other words, the user can test any text inputs, ranging from plain sentences, sarcasm, and ambiguous terms like "meh". This function will help facilitate the link from academic research to real-world implementation, and finally demonstrate development of a sentiment analysis systems that can be utilized in the real-world.

4.8 Evaluation

The research comprehensively addressed all established research objectives through systematic research, achieved all stated research objectives through diligent empirical assessment and practical application. The comparative analysis made transparent the trade-off in performance/efficiency characteristics of the BERT and DistilBERT architectures. The models present very high performance when classifying sentiment but have markedly different computing characteristics that lend themselves to different styles of deployment. The implementation of confidence-based neutral classification with a threshold of 70% was specifically designed to address the concerns about ambiguous sentiment expression and also provided some nuanced classification that simple binary systems cannot. The research also demonstrated that BERT and DistilBERT fine-tuning methods enables both architectures to effectively evaluate complicated linguistic constructs such as sarcasm without needing specialized detection architecture. The systematic evaluation process demonstrated the basis for trustworthy conclusions about architectural characteristics and illustrated meaningful practical deployment approaches through a multi-stage interactive model selection to deploy, defining real-world applicability and usability considerations beyond mere accuracy metrics for production sentiment analysis systems.

5. Conclusion and Directions for Future Work

5.1 Overview

This research enabled to present a comparative evaluation on intent classification based on transformers using meaningful, more fully implemented, additional functionality, expanded functionality relative to confidence-based neutral classification and sarcasm detection. The results indicate that both transformer architectures are able to perform sentiments classification, with equal performance. Given the current discoveries, the interactive model selection system will allow an organisation to choose between the two architectures based upon outcomes. For instance, BERT will outperform DistilBERT for maximum accuracy but if deployment is optimised, DistilBERT will be of value. To summarise from the comparative evaluations, DistilBERT retains 99.76% of the accuracy of BERT in classification accuracy. Also, DistilBERT indicated significant computing; for example, 47% reduction in training time, 90% in processing time. With these positive metrics, this project supports the concept of knowledge distillation as a usable technology for the implementation of transformer models in resource constrained environments. Moreover, the use of extra functionality with the confidence-based neutral classification system at a threshold of 70% retains function to remove the restrictions associated with conventional binary sentiment based classification systems; and both models were competent with sarcasm using ordinary fine tuning methods and no need for a special architecture.

5.2 Practical Applications

The study results have achieved meaningful practical relevance; across several industries needing automated sentiment analysis functionality. E-commerce companies can use the comparative framework to monitor the sentiment of customer reviews, by choosing BERT when accuracy is vital for serious comments and DistilBERT when feedback is processing high volumes of reviews in real time. Provided the confidence-based neutral classification provides insight into customer sentiments that are uncertain (needing human attention), and can capture sarcasm or complex customer sentiment, automated response functionality has a new dimension. Changes enabling DistilBERT to process feedback 90% faster than BERT, it offers the opportunity for real time sentiment tracking of social media campaigns, customer service interactions or market research projects. The interactive model selection mechanism will support organisations to make informed decisions about how to deploy them considering any computational challenges alongside accuracy. This provides a way of operationalising the transition from traditional models of sentiment analysis to transformer approaches; allowing innovation benefits in efficiency of deployment to be realised in situations where resource constraints necessitate reduced analysis capability. Effectively making advanced sentiment analysis functionality available for small organisations with limited computational infrastructure.

5.3 Limitations and Future Enhancements

This study has limitations, implications which will impact generalisability. The data source was restricted to English language Amazon reviews which will ultimately limit generalisability for sentiment analysis on multi-lingual populations and in domains outside of e-commerce. The detection of neutrality was set at an empirical confidence threshold of 70% and was not tested with systematic tuning which means that it may not be the optimum threshold for detection across domains. The evaluation only reviewed sentiment in customer reviews and has not reviewed performance on any other text type, i.e. social media posts and formal documents. The research was conducted in a controlled experimental research context where variability of real-world and user input challenges/quality data issues may not have been realised.

5.4 Future Research Directions

There are many paths to consider guiding the development of the research framework in multiple ways. Assessing multi-lingual sentiment analysis would enhance generalisability with the difference in languages. Dynamic confidence thresholds could enable the model's classification thresholds to be dynamic and context dependent to the text, and could be developed to further improve neutral detection across contexts. Future work could also assess more robust comparisons of models by assessing more transformer variations to validate baseline results. Testing domain adaptation would be an assessment of the models performance on wider text types than customer reviews alone. Finally, collaborating to include live data sources would enable live sentiment tracking for social media partners and corporate customer service tools.

ACKNOWLEDGEMENTS

I would like to first acknowledge my project supervisor Barron Jason who provided an immense amount of help and support throughout this research project and whose knowledge helped assist in the completion of this work. I would also like to acknowledge and give thanks to the developers of the Hugging Face Transformers library and the many developers of the Amazon Review Polarity Dataset which both contributed their work available to the research community. Finally, I would like to acknowledge the opensource development community (especially the developers of PyTorch and scikit-learn) for their tools to assist in implementing this research.

REFERENCES

[1] Amal M. Areshey, Sherin S. Mathew and M. Supriya. 2024. Transformer models for sentiment classification of COVID-19-related tweets. *Natural Language Engineering* 30, 2 (2024), 265-302.

[2] Tanvi Dadu and Kartikey Pant. 2020. Sarcasm Detection using Context Separators in Online Discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 51–55, Online. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.figlang-1.6

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[4] Clayton H.E. Gilbert and Eric Hutto. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the eighth international AAAI conference on weblogs and social media* (Vol. 8, No. 1). 216-225.

[5] Albert Gu, Karan Goel and Christopher Re. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

[6] Ansu Ann Joshy and Sheeba Sundar. 2024. Sentiment analysis of English-translated Tamil literature using pre-trained transformer models. *Expert Systems with Applications* 238, p.122200.

[7] Salman Khan, Syed Hakak, N. Deepa, Keshav Dev and Gyanendra Reddy. 2024. Transformer attention-based approach for sarcasm detection in social media. *Journal of Big Data* 11, 1, p.45.

[8] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 2019. RoBERTa: A robustly optimised BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[10] Usman Naseem, Imran Razzak, Katarzyna Musial and Muhammad Imran. 2020. Transformer based deep intelligent contextual embedding for Twitter sentiment analysis. *Future Generation Computer Systems* 113, 58-69.

[11] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing* (Vol. 10). 79-86.

[12] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

[13] Anna Rogers, Olga Kovaleva and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics* 8, 842-866.

[14] Konstantinos I. Roumeliotis, Nikolaos D. Tselikas and Dimitrios K. Nasiopoulos. 2024. A comparative study of pre-trained language models for text classification tasks. *Information* 15, 2, p.102.

[15] Victor Sanh, Lysandre Debut, Julien Chaumond and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[16] Himani Srivastava, Vaibhav Varshney, Surabhi Kumari and Saurabh Srivastava. 2020. A Novel Hierarchical BERT Architecture for Sarcasm Detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 93–97, Online. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.figlang-1.14

[17] Kai Sheng Tai, Richard Socher and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

[18] Enrica Tian, Nan Zhang and Peixin Chen. 2023. Dynamic Routing Transformer Network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 2543-2556.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998-6008.

[20] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

[21] Juliann Zhou. 2023. An Evaluation of State-of-the-Art Large Language Models for Sarcasm Detection. *arXiv preprint arXiv:2312.03706*. DOI: https://doi.org/10.48550/arXiv.2312.03706