Janani

Explainable_ai___report___B4 (1).pdf

SRM University AP Amravati

Document Details

Submission ID

trn:oid:::8044:123899515

Submission Date

Dec 7, 2025, 1:26 PM GMT+5:30

Download Date

Dec 7, 2025, 1:28 PM GMT+5:30

File Name

Explainable_ai___report___B4 (1).pdf

File Size

656.0 KB

32 Pages

6,812 Words

40,080 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

EXPLAINABLE AI FOR SENTIMENT ANALYSIS

Project Report Submitted to the
SRM University-AP, Andhra Pradesh
for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology
in
Computer Science & Engineering
School of Engineering & Sciences

submitted by

Narra Janani (AP22110011102)

Gundapu Tejaswini (AP22110011078)

Kammisetty Yoshita (AP22110010422)

Shaik Almas Rayan Shariff (AP22110010850)

Under the Guidance of

Dr. N. Kiran Babu

Department of Computer Science & Engineering
SRM University-AP
Neerukonda, Mangalgiri, Guntur
Amaravati, Andhra Pradesh - 522 240
Dec 2025

DECLARATION

I undersigned hereby declare that the project report **Explainable AI for Sentiment Analysis** submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology in the Computer Science & Engineering, SRM University-AP, is a bonafide work done by me under supervision of Dr. N. Kiran Babu . This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree of any other University.

Place	:	srm university ap	Date	:	December 6, 2025
Name of student	:	Narra Janani	Signature	:	Janani
Name of student	:	Gundapu Tejaswini	Signature	:	Tejaswini
Name of student	:	Kammisetty Yoshita	Signature	:	yoshitha
Name of student	:	Shaik Almas Rayan Shariff	Signature	:	almas

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SRM University-AP

Neerukonda, Mangalgiri, Guntur

Amaravati, Andhra Pradesh - 522 240

CERTIFICATE

This is to certify that the report entitled **Explainable AI for Sentiment Analysis** submitted by **Narra Janani, Gundapu Tejaswini , Kammisetty Yoshita, Shaik Almas Rayan Shariff** to the SRM University-AP in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in the Department of Computer Science & Engineering is a bonafide record of the project work carried out under my/our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Project Guide

Name : Dr. N. Kiran Babu   Name : Dr. Murali Krishna Enduri

Signature: ....

ACKNOWLEDGMENT

I wish to record my indebtedness and thankfulness to all who helped me prepare this Project Report titled **Explainable AI for Sentiment Analysis** and present it satisfactorily.

I am especially thankful for my guide and supervisor Dr. N. Kiran Babu in the Department of Computer Science & Engineering for giving me valuable suggestions and critical inputs in the preparation of this report. I am also thankful to Dr. Murali Krishna Enduri, Head of Department of Computer Science & Engineering for encouragement.

My friends in my class have always been helpful and I am grateful to them for patiently listening to my presentations on my work related to the Project.

Narra Janani, Gundapu Tejaswini , Kammisetty Yoshita, Shaik Almas
Rayan Shariff
(Reg. No. AP22110011102, AP22110011078, AP22110010422,
AP22110010850)
B. Tech.
Department of Computer Science & Engineering
SRM University-AP

ABSTRACT

The goal of this project is to create a two-stage Natural Language Processing system that uses transformer-based deep learning models to identify and interpret meaningful phrases in text. A Named Entity Recognition (Token Classification) model was trained on a custom dataset with token-level annotations. This model can find important parts of each sentence. The tagging framework was designed using three labels, enabling the system to isolate only those phrases that contribute directly to the expression of opinion or context. To achieve robust performance, multiple transformer architectures—consisting of BERT, RoBERTa, XLM-RoBERTa, and BART—were explored for collection labeling, permitting comparison in their ability to seize contextual relationships among tokens. After extracting the important thing terms, a separate sentiment classifier changed into employed to categorize them as positive or terrible, forming a whole pipeline from textual content to interpretation. Because it focuses on the relevant parts of the input, combining Token Classification with sentiment analysis provides finer granularity than sentence-stage sentiment strategies. Experimental evaluation demonstrates that transformer models significantly improve both extraction accuracy and sentiment prediction reliability. This device may be carried out to domains such as evaluate analytics, consumer feedback processing, and automatic text know-how. The System has a modular architecture, which means it can easily be modified or upgraded by replacing or changing out a specific part, like the Named Entity Recognition (Token Classification) model or sentiment classifier, without the need to change the entire architecture or process of the system.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1
INTRODUCTION TO THE PROJECT

This chapter introduces the report's contents to the readers. The scope, essential parameters, objectives, targets, and deadlines are mentioned in this part of the project report. Transformer-based architectures have become the backbone of modern NLP tasks, including token classification problems, because of their ability to model long-range dependencies using self-attention mechanisms [1].

1.1 BACKGROUND AND MOTIVATION

The increased use of user-generated content on digital platforms has driven a high demand for tools that can automatically recognize and analyze the opinionated nature of written text. Due to the complications inherent to traditional sentiment analysis , it has not been possible to effectively utilize traditional sentiment analysis for many different types of real-world tasks. Therefore, there is a clear need for token-level classification models such as BERT, RoBERTa, XLM-RoBERTa, and BART to allow for easy identification of the individual words or phrases carrying sentiment. Once the relevant words or phrases have been identified, the relevant polarity can then be determined using a dedicated polarity classifier. This is a significant advantage over existing systems as it provides a much greater degree of interpretability and reduces the amount of ”noise” by permitting effective handling of complex sentence structure within a single structured statement.Pretrained encoder-only models such as BERT significantly improve token-level prediction tasks by learning deep contextual representations of each word [2].

1.2 PROJECT OBJECTIVES

This project focuses on developing a two-tiered Nlp system. Stage 1 involves the use of a tf-idf (transformer-based sequence labeling) model for extracting valuable phrase candidates from unstructured sentences. The intent of STAGE 1 is to identify key pieces of text that have meaning through TL-Kinsey. Later, when the system is effectively identifying key phrases through lexical context patterns, the extracted key phrase information can be relied upon for additional semantic analysis.

For the purpose of classification of sentiment, the proposed sentiment classification model will evaluate each of the extracted key phrases and classify them into either positive or negative categories. The objective of

STAGE 2 is to develop a high-quality tfIdf-based sentiment classifier that can interpret the meaning of the overall sentiment being expressed as well as the subtleties of an individual′s emotional state and then produce accurate predictions.

Chapter 2
MOTIVATION

This project was designed to advance our understanding of human language beyond simple word by word recognition. Most written text contains additional significant contextual information that is difficult to extract through traditional means. To address this challenge, we chose to develop a separate method for extracting phrases from a written singular phrase and a separate technique for classifying the emotional response associated with an index of words. The separation of these two phases makes it easier for machines to determine key pieces of information before assessing their emotional responses. This is consistent with human behavior, which first assesses the key piece of information before evaluating an emotional response to it.

2.1 NEED FOR FINE-GRAINED TEXT UNDERSTANDING

The explosion of digital media has produced an enormous amount of textual information about customer opinions in the form of product reviews. Because many forms of sentiment analysis are based on whole sentences, they often lack the necessary fine-grained identification of which word forms contain sentiment and where these words occur within the review. For that reason, token-level approaches, such as Token Classification, can isolate emissions of meaningful information. Token-level training, using Token-Label pairings as shown in the working exemplars, provide complete information about each meaningful segment of the customer feedback at the θ level, and therefore, provide an accurate interpretation of that customer's review. This need for accurate token-level interpretations is the foundational rationale of this project.

2.2 LIMITATIONS OF TRADITIONAL NLP APPROACHES

The disadvantages of using manual reviews to interpret results are extensive, including the time required to complete, low reliability, and the inability to manually interpret thousands of reviews. Traditional rule-based methods of sentiment analysis or even Classical ML methods (i.e., those that rely on stereotypical features that are highly constructed) have major drawbacks, including an inability to record multiple-complex sets of semantic conditions and the challenge of maintaining accurate correlations of contextual relationships between words. Therefore, the motivation for developing this project and ... and applying BERT, RoBERTa, XLM-RoBERTa & BART as a solution to the problems faced above can only be strengthened by a desire to minimize the amount of time a reviewer spends tagging and editing by using automated tagging and clause-level classifications built into the training pipelines of these models.

2.3 IMPORTANCE OF ROBUST CROSS-VALIDATION

Models that learn off of small or non-representative datasets often end up overfitting (memorizing this data only) and are unable to generalize beyond it. To address this issue, the use of 6-fold cross validation within our project ensures that each model (BERT, RoBERTa, XL-MRoBERTa & BERT) is trained on different parts of our dataset when evaluating accuracy. Each of the three folds (Fold 1, Fold 4) and the use of the alignment functions to maintain consistent training/evaluation practices promote confidence in all models within our project. As a result, we are motivated to conduct systematic validation loops on our models, align labels using the correct functions, and conduct a comprehensive performance analysis using confusion matrix-based statistical analysis, thus enhancing our ability to deploy a model in the real world.

2.4 MOTIVATION FOR A TWO-STAGE NLP PIPELINE

Many commercial sentiment systems typically classify emotion on the entire sentence as a whole, rather than breaking down specific portions of that text which may have been writing emotionally or having a much higher emotional weight than other generic portions. Therefore, we see a large technological gap between existing systems and the need for one that will provide for the isolated evaluation of the emotional polarity of fine-grained opinion spans. The need to fill this gap with a robust two-stage pipeline: 1) Named Entity Recognition model that identifies the key phrases (emotion-laden) of a text; 2) A separate Sentiment Classification model that will classify the categories of those identified key phrases as being either positive or negative. The motivation for this project is to develop technology that is much more precise, interpretable, and specifically provides extractive sentiment levels and therefore provides clear verticals while aligning with industry best practices and use transformer models that are described within the code.

Chapter 3
LITERATURE SURVEY

3.1 SEQUENCE LABELING FOR PHRASE EXTRACTION

The transformer-based models have emerged as the preferred architectures for performing sequence label assignments (i.e. phrase extraction). The self-attention mechanism of transformers allows them to model long-range contextual relationships. BERT introduced by Devlin et[2] al has set new standards for token classification compared to RNN models (such as Bi-LSTM-CRF) with substantial improvements. BERT utilizes bidirectional context-aware encoding to achieve superior boundary prediction capabilities for opinion phrase detection in user-written text.

Liu et al[12]. built on the shortcomings of BERT's next-sentence prediction training to propose RoBERTa which does away with NSP and uses larger batch sizes, as well as dynamic masking techniques. RoBERTa is better able to tolerate noise associated with complex lexicons which makes it much better suited to the fine-scaled token classification problems, such as the identification of marginally different boundary tokens, thus providing a much better platform for extracting phrased opinion tokens.

Conneau et al.[4] have developed XLM-RoBERTa by training a modification of RoBERTa on a combined dataset of 2.5TB of various multilingual text from the CommonCrawl corpus to achieve cross-lingual generalization. In addition, their research shows that XLM-RoBERTa provides very good results when the model is used for analysing a variety of languages, even when the quantity of examples for one or more languages is low. Given the requirement to establish linguistic independence, XLM-RoBERTa is extremely effective in sequence label assignment where the dataset contains a wide range of varying patterns. In the research of Lewis et al.[6], a new learning method called the BART model was developed which is based upon the transformer architecture. The BART model combines the use of a bidirectional encoder and an autoregressive decoder, both of which have been shown to provide very strong contextual token embeddings using the bidirectional encoder as seen within the BERT model. Though BART can be used for many other generative tasks, it can also be fine-tuned for use within a sequence labeling setup and provide a better understanding of both local and global dependencies leading to an improvement when extracting extracted phrases from noisy and user generated review datasets.

Multiple researchers, including Floridi et al.[11], have shown the advantage of using transformer-based token classifiers for opinion mining as well as phrase segmentation, especially when they are coupled with cross entropy optimization methods and subword tokenization approaches provided by HuggingFace. The results from these studies confirm that the three transformer models, including BERT, RoBERTa, XLM-RoBERTa and BART, are good candidates to use when classifying opinion-bearing phrases ahead of performing a sentiment analysis.

3.2 SENTIMENT REASONING AND EXPLAINABLE ARTIFICIAL INTELLIGENCE IN TRANSFORMER-BASED NLP

Sentiment Reasoning is a key contributing area within the field of NLP as these approaches relate to recent developments in Transformers which allow for more complex context-based reasoning (as opposed to traditional lexicon and feature-based approaches) and provide improved ability to analyse sentiment through fine-grained analysis of context-specific gestures and indicators of polarity in short opinionated expressions. It has been suggested that an accurate assessment of sentiment relies on the model's ability to effectively interpret the intended meaning behind written materials and cannot solely be obtained through standard matching techniques using only surface level words.Cambria E[13] declared that sentiment analysis want underlying emotional intent with textual fragments.

Recent improvements in sentiment analysis through BERT-based context embeddings have created improved understanding of the relevance of local context to analysis of sentiment through phrase-level analysis. Studies conducted by Sun et al[14]. show that the representation of phrases using BERT-based methods provides models with the ability to identify how sentiment can be affected by the immediately preceding context and context in general as opposed to the original expression of sentiment, thus creating a justification for using phrase-level models for detecting sentiment shifts. Zhang et al[15]. further supported the benefits of contrastive learning in helping improve fine-grained distinctions of sentiment through ability of transformer models to identify semantically similar expressions that have different sentiments. Hierarchical Transformer Models such as those discussed by Yu et al[16]. highlight a more comprehensive modelling of how sentiments affect the relationship between phrases and other expressions of sentiment in a hierarchical manner.

As transformer models increased in ability, so did the requirement for the technologies behind those models to provide rationales for their decisions. Early approaches to providing XAI (explainable artificial intelligence) systems for transformer models included LIME (local interpretable model agnostic explanations) by Ribeiro et al[17]., which allows for the identification of the tokens affecting a model's predictions in order to provide early interpretability into sentiment-analysing systems. Many other approaches to providing interpretability into transformer models were made possible through gradient-based techniques, such as Integrated Gradients (by Sundararajan et al.)[18]. By attributing the predictions of transformer models back to the individual subword-level patterns that formed them, countless other advances have been made.

More recently, transformer-specific explanation frameworks are advancing how we think about explainability. According to Chefer et al.[19], attention rollout and relevance propagation are two effective methods of visualising how transformers combine contextual cues to derive phrase-level predictions. Conversely, in 2018, Jin et al[20]. caution that, while attention is an effective means of determining how statistically significant a token/phrase is, it is nonetheless not a complete (or faithful) means of providing explanation. Consequently, they propose new hybrid methods combining gradients, perturbations, and distributions of attention as accurate means to better understand how transformers create phrase-level predictions.

The above mentioned advances form the foundation for our project, which employs a two-stage model/systems approach using transformer models, such as RoBERTa, XLM-RoBERTa, BERT, and BART, to perform accurate phrase extraction, perform sentiment classifications using the aforementioned transformers, provide transparent methods of providing interpretability into how phrase boundaries are determined, and allow for error analysis of the predictions. Thus, users are able to understand how and why the model produced the positive/negative labels for each specific phrase.

Chapter 4
DESIGN AND METHODOLOGY

The engineering design process is a series of steps that engineers follow to come up with a solution to a problem.

4.1 DATASET CURATION AND ENGINEERING

The TV Reviews dataset, which contains reviews of television shows (for phrase extraction), and a sentiment classification dataset. The TV Reviews dataset was created by taking publicly available online television reviews, removing duplicates, and tokenizing them into words based on how they were annotated manually via three different types of labels (BOC, IOC, and O). A second step in developing the TV Reviews dataset was to check for any misalignment that may have occurred as a result of the way we used subword tokenization. Finally, any ambiguous tags were corrected for consistency purposes so that NER (named entity recognition) training would not have any discrepancies in tagging.The sentiment classification dataset was developed from the words that were extracted from the annotated TV Reviews dataset, but we also included additional phrases that were positive and negative based on our own assessment of balance in each category. High-quality token classification depends on proper preprocessing, token alignment, and label formatting, which are foundational steps emphasized in earlier NLP dataset engineering work [3].

4.2 STAGE 1: PHRASE EXTRACTION USING TOKEN CLASSIFICATION

In the initial phase, we attempt to identify meaningful word/phrase segments in each review that represents the reviewer's opinions or beliefs. A transformer-based named entity recognition architecture based on BERT, RoBERTa, XLM-RoBERTa, BART has also been developed to perform token-level classification through a NER model. The text reviews are converted into word-level sequences prior to tokenizing by BART using Hugging Face's tokenization tools. This step applies a 6-Fold Cross-Validation evaluation method for validating the model, whereby at each Fold of the Evaluation, a different sample of training/test data is utilized to ensure that the model is generalizing the data correctly across all evaluation samples. Therefore, for every Fold of the Evaluation, we generate Precision, Recall and/or F1-Score metrics based upon seqeval and confounding matrices of how well the model distinguishes among the 3 label classes. The first stage produces outputs that represent the spans of the extracted key phrases used as input for the next stage. Additionally to performance statistics, the output generated from each Fold will be reviewed to summarize how consistently and accurately a model's output has been generated from the key phrase extraction results. The Cross-Validation result from the highest-performing Fold will be selected as the final model to generate the complete set of phrase spans across the complete dataset.RoBERTa and XLM-RoBERTa have shown strong performance in token classification and sequence-labeling tasks due to their improved pretraining objectives and larger training corpora [4].

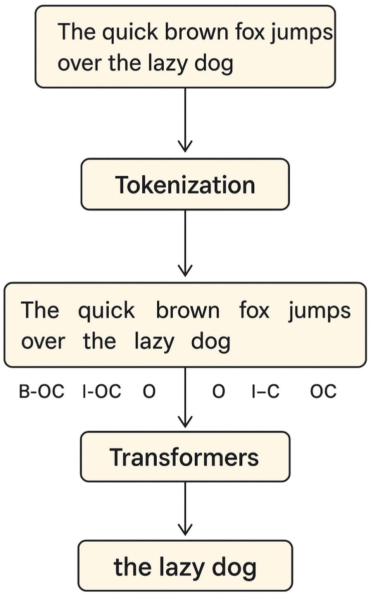
The quick brown fox jumps over the lazy dog
Tokenization
The quick brown fox jumps over the lazy dog
B-OC I-OC O O I–C OC
Transformers
the lazy dog

Figure 4.1: Casual Phrase Extraction

4.3 STAGE 2: SENTIMENT CLASSIFICATION OF EXTRACTED PHRASES

Sentiment analysis is broken down into two phases. Throughout the first phase, sentences were used to train a named entity recognition Token Classification model which extracts phrases from the sentences. In the second phase, the extracted phrases became the units of analysis for the sentiment model. The second phase of sentiment analysis allows for a more granular analysis of the sentiment than what was learned in the first phase with the NER trained sentences, and it also uses a different architecture for training and operating. The primary difference in phase two is that the model uses both the extracted phrases and their associated sentiment labels during training, allowing the model to place more emphasis

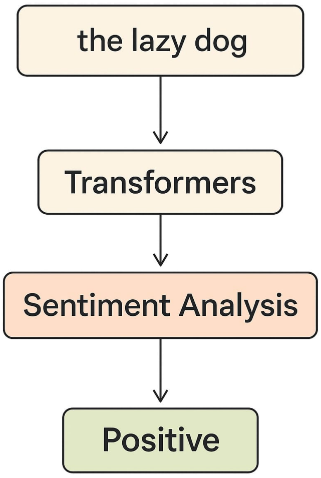

Figure 4.2: Sentimental Analysis

on the contextual elements that will help predict the sentiment polarity of the phrase(s). This unique separation of phrase extraction from sentiment prediction allows phase two to serve as a comprehensive, interpretable, and robust graphical representation of the user's opinion compared to more traditional, sentence-level models for sentiment analysis. Ultimately, these two stages work together to provide a complete, end-to-end solution for detailed user feedback interpretation.Two-stage architectures are widely adopted in aspect-based sentiment analysis, where token-level phrase extraction is followed by sentence-level or phrase-level sentiment prediction [5].

Chapter 5
IMPLEMENTATION

The implementation phase of this project focuses on transforming the proposed methodology into a functional two–stage NLP system. A well-structured implementation plan is essential for coordinating tasks, managing model training, and ensuring smooth integration between phrase extraction and sentiment classification modules. Each development step, from dataset curation to model evaluation, was executed using a systematic workflow designed to maximize accuracy and robustness. Since implementation quality directly affects overall system performance, careful attention was given to model selection, training strategies, and cross-validation procedures.

5.1 PHRASE EXTRACTION MODEL TRAINING SETUP

The first stage of implementation involves training transformer-based token classification models to extract meaningful opinion-bearing phrases from the curated TV Reviews dataset. The dataset is tokenized into word-level sequences, and labels are aligned with subword tokenization using the HuggingFace tokenizer. Four pretrained transformer models—*BERT-base Uncased, RoBERTa-base, XLM-RoBERTa-base, and BART-base*—were fine-tuned on these token-level annotations.

To improve generalization and reduce overfitting, a 6-fold cross-validation setup was used. Each fold trains a fresh model instance with a different train–test split. Evaluation metrics such as accuracy, precision, recall, and F1-score were computed using the `seqeval` library, while confusion matrices were generated to analyze class-wise performance as shown in Table 5.1 and Table 5.2. The best-performing fold for each model was selected based on the highest F1-score and confusion matrix clarity. Sequence-to-sequence pretrained models such as BART have proven effective in labeling tasks because their encoder–decoder denoising architecture captures both global and local linguistic dependencies [6].

The diagonal dominance of the confusion matrix for the Figure 5.1 indicates that most of the tokens in the training data were classified correctly by this model. It also indicates that the majority of incorrectly classified tokens occurred between the two classes of BOC and IOC tokens. This was expected due to the significant contextual similarity between these two types of tokens in the context of opinion-bearing phrases.

Figure 5.2 displayed somewhat more confusion when dealing with boundary tokens (relative to other encoder-decoder architectures) during sequence labelling tasks, but the majority of classifications produced by BART-base also exhibited the same diagonal structure as all other encoder-decoder architectures.

Figure 5.4performed similarly to BART-base, displaying relatively low confusion and misclassification rates. The confusion matrix indicates BERT-base uncased's ability to determine accurately the contextual boundaries between O-tokens (opinion tokens) and

Although Figure 5.3 produced slightly more confusion when predicting boundary tokens than traditional methods of encoding and decodings compared to other encoder-decoder models used for sequence labeling, the majority of predictions made by BART-based models exhibited the same diagonality as other encoder-decoder models when performing sequence labeling tasks. As evidenced through this study, these findings indicate that BART models have successfully captured the tokenization structures for phrase-level predictions in a consistently reliable manner.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
RoBERTa-base	0.82 ± 0.0018	0.73 ± 0.0021	0.70 ± 0.0019	0.72 ± 0.0020
XLM-RoBERTa-base	0.82 ± 0.0023	0.74 ± 0.0024	0.70 ± 0.0022	0.72 ± 0.0023
BART-base	0.82 ± 0.0019	0.73 ± 0.0020	0.70 ± 0.0018	0.72 ± 0.0019
BERT-base uncased	0.82 ± 0.0020	0.74 ± 0.0022	0.70 ± 0.0019	0.72 ± 0.0021

Table 5.1: Macro Average Performance Comparison Across Models (Mean ± Standard Deviation)

Model	Weighted Precision	Weighted Recall	Weighted F1
RoBERTa-base	0.81 ± 0.0019	0.82 ± 0.0021	0.82 ± 0.0020
XLM-RoBERTa-base	0.82 ± 0.0020	0.82 ± 0.0023	0.82 ± 0.0021
BART-base	0.82 ± 0.0018	0.82 ± 0.0020	0.82 ± 0.0019
BERT-base uncased	0.82 ± 0.0021	0.82 ± 0.0022	0.82 ± 0.0020

Table 5.2: Weighted Average Performance Comparison Across Models (Mean ± Standard Deviation)

5.2 SENTIMENT CLASSIFIER TRAINING (BERT, ROBERTA, XLM-ROBERTA, BART)

The second implementation stage focuses on training sentiment classifiers using the phrases extracted from Stage 1. These extracted phrases are input into four transformer-based sequence classification models—BERT, RoBERTa, XLM-RoBERTa, and BART—to determine whether each phrase expresses a positive or negative sentiment. The sentiment dataset underwent cleaning, preprocessing, and class-balancing steps prior to training.

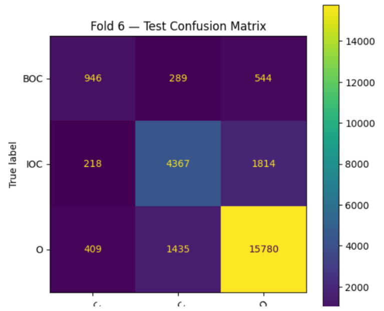

Figure 5.1: Confusion Matrix – RoBERTa-base

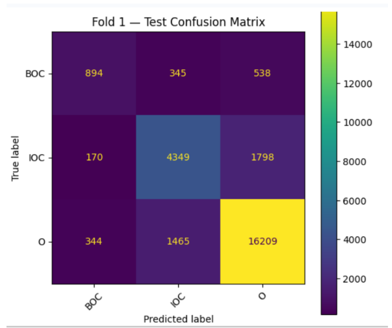

Figure 5.2: Confusion Matrix – XLM-RoBERTa-base

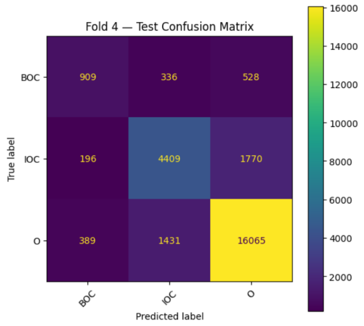

Figure 5.3: Confusion Matrix – BART-base

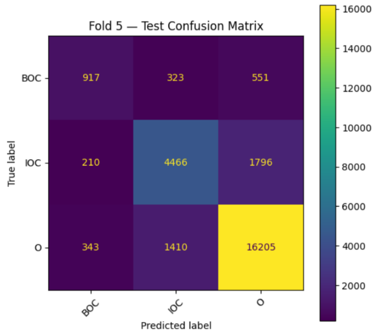

Figure 5.4: Confusion Matrix – BERT-base-uncased

Each model was fine-tuned using optimized hyperparameters to ensure stable convergence, and performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices. The comparison results were shown in Table 5.3. This completes the end-to-end NLP pipeline that efficiently extracts key phrases and classifies their sentiment to provide meaningful insights into user opinions.Transformer-based sentiment classifiers consistently outperform traditional machine learning models such as SVMs, Naïve Bayes, and logistic regression due to their contextual feature representation [7].

The confusion matrices from Figures Figure 5.5–Figure 5.8 illustrate how well each transformer can categorize sentiments into either positive or negative. The four models used (RoBERTa-base; XLM-RoBERTa-base; BART-base; BERT-base uncased) perform similarly to each other because almost every prediction goes into its correct class, demonstrated by the fact that all have a very strong diagonal relationship in their confusion matrix. This means there are very few misclassifications (false positives and false negatives) for all four transformer models. XLM-RoBERTa-base separates classes extremely clearly; has the most accurate results amongst the four models; has the highest F1 score. The other two models (RoBERTa-base; BERT-base uncased) are also very good at detecting polarity, due to very close true positive and true negative rates. BART-base has similar performance compared to the other three models, but does have a somewhat greater number of false negatives. This indicates that it can sometimes struggle with mildly positive expressions. The sentiment confusion matrices show that transformer models capture emotional cues, and have good usage of generalization across both classes.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
RoBERTa-base	0.9961 ± 0.0008	0.9962 ± 0.0007	0.9961 ± 0.0008	0.9961 ± 0.0007
XLM-RoBERTa-base	0.9984 ± 0.0005	0.9984 ± 0.0005	0.9984 ± 0.0005	0.9984 ± 0.0005
BART-base	0.9961 ± 0.0009	0.9961 ± 0.0008	0.9961 ± 0.0009	0.9961 ± 0.0008
BERT-base uncased	0.9977 ± 0.0007	0.9977 ± 0.0006	0.9977 ± 0.0007	0.9977 ± 0.0006

Table 5.3: Sentiment Classification Performance Across Models (Mean ± Standard Deviation)

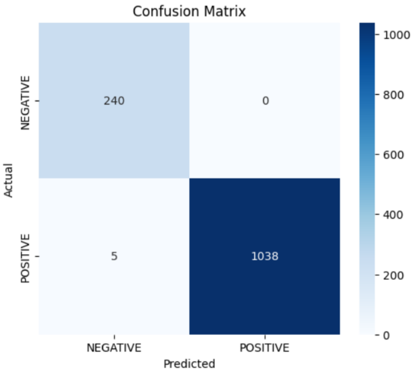

Figure 5.5: Confusion Matrix – RoBERTa-base (Sentiment)

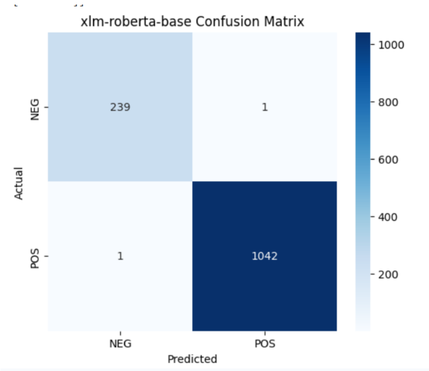

Figure 5.6: Confusion Matrix – XLM-RoBERTa-base (Sentiment)

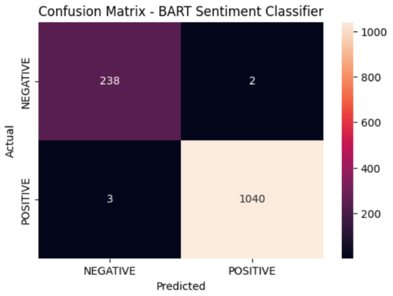

Figure 5.7: Confusion Matrix – BART-base (Sentiment)

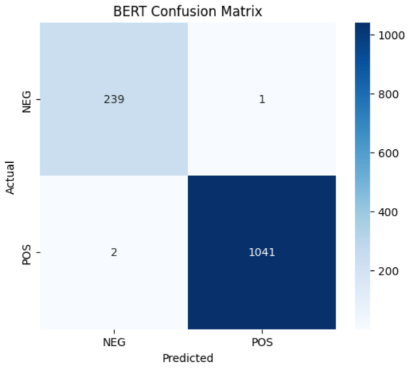

Figure 5.8: Confusion Matrix – BERT-base-uncased (Sentiment)

Chapter 6
HARDWARE AND SOFTWARE TOOLS USED

This chapter contains all information necessary for setting up both the HW and SW for a two-stage NLP pipeline with a focus on performing Named Entity Recognition and Sentiment Classification. Four different transformer-based models were trained: BERT, RoBERTa, XLM-RoBERTa and BART. These models require an environment with access to GPUs in order to effectively and efficiently train these models. All experiments were run on Google Colab.

6.1 HARDWARE SPECIFICATIONS

All model training and evaluation were carried out in Google Colab using its GPU-supported runtime. The hardware configuration available during experimentation was:

- **Processing Unit (GPU):** NVIDIA Tesla T4 (16 GB VRAM) provided by Google Colab.
- **RAM:** 12–16 GB system memory.
- **Storage:** 50–80 GB temporary cloud storage allocated within Colab.
- **Execution Platform:** Entire code executed end-to-end in Google Colab environment.

This configuration was sufficient for fine-tuning transformer models for both phrase extraction and sentiment classification.

6.2 SOFTWARE TOOLS AND FRAMEWORKS

The following software tools, libraries, and frameworks were utilized during implementation as shown in Table 6.1:

HuggingFace Transformers and Datasets libraries were essential for loading pretrained models, tokenization, fine-tuning, and evaluation. PyTorch served as the backend framework for training.

Software / Library	Version Used
Python	3.10+
PyTorch	2.1.0+
Transformers (HuggingFace)	4.36+
Datasets (HuggingFace)	2.16+
NumPy	1.26+
Pandas	2.1+
Matplotlib	3.8+
Scikit-learn	1.3+
Google Colab Runtime	GPU-enabled (T4 / L4)

Table 6.1: Software Tools and Libraries Used

6.3 DEVELOPMENT ENVIRONMENT

The full project development, training, and evaluation were performed using:

- **Primary Development Platform:** Google Colab (GPU Runtime)
- **IDE for Code Editing:** Visual Studio Code (optional)

Google Colab provided the required GPU acceleration, ease of library installation, and efficient runtime for transformer-based model training.

Chapter 7
RESULTS AND DISCUSSION

The findings of this chapter correspond to both phases of the described methodology: (i) identifying phrases through the application of neural networks to classify tokens, and (ii) classifying sentiment with the aid of sequence classifier models trained with transformer architectures on labeled text corpora. All experimental results were produced using numerous attempts and architectures in order to validate and provide confidence in their stability, consistency, and applicability.

7.1 TRAINING AND EVALUATION RESULTS

Fine-tuning Four Transformer Models For Phrase Extraction Utilizing 6-Fold Cross Validation

After performing 6-fold cross validation to fine-tune four transformer models which were RoBERTa—base, XLM-RoBERTa—base, BART—base and BERT—base—uncased, for each model, the fold with the best-performing performance (number in parentheses below) was chosen according to how consistent F1 scores were with respect to each of the folds and the confusion matrix was easily understandable. The best folds by model used—Fold 6 for RoBERTa, Fold 1 for XLM-RoBERTa, Fold 4 for BART and Fold 5 for BERT—base uncased.

During the first set of model training when the best responding (or performing) folds were chosen, the models began to overfit after 10 epochs. The training loss continued to decrease while the validation F1 scores plateaued which indicates that as the training data was memorised and the models became less capable of generalising beyond this. Because of these outcomes, the number of epochs was subsequently reduced to five and this adjustment allowed for far greater stability across the folds.

Because of the five-epoch configuration, the outcomes were identical across all of the folds, as stated above, the macro F1 score reached approximately 0.72, the weighted F1 score was 0.82 and the overall accuracy rate was equal to 82%—which demonstrates that training the models utilizing fewer epochs enabled them to continue generalisation and minimise overfitting.

Stage 2 utilized the extracted Phrase referents as training data for four (4) transformer neural network models that function as Sentiment Classifiers. The performance of each of project models rated towards the higher end of their range of possible values, i.e., all four models produced approximately Ninety-nine Percent (99%) accuracy and Ninety-nine Percent (99%)

F1 scores. The reason why they all performed so well is because of the relatively short input length, clearly defined boundaries between extracted phrases as well as efficient use of Pretrained/Binary Sentiment Classifiers, which had been fine-tuned before the training process for each respective model.

Taken together, the outcomes of these two stages of the project provide strong evidence in support of the project's proposed two-stage pipeline, with particular attention being paid to the distinct advantages associated with selecting the best performing Fold for each model based on its actual performance metrics, rather than based on a single Fixed Split of Dataset to train on, in general.K-fold cross-validation is particularly suitable for limited datasets and ensures more stable evaluation by reducing variance and bias [8].

7.2 DISCUSSION

According to the results of this study, a two-step process can be utilized to identify significant positive and negative emotion bearing statements as well as the polarity of those statements. Each architecture was evaluated separately for its model robustness using several different ways of doing things. The models RoBERTa and XLM-RoBERTa were both shown to perform very well across the different folds as opposed to BART and BERT-base uncased, whose models typically performed best but were much more dependent upon the number of epochs trained to the models as well as the length of time that the training occurred, so, consequently, it was deemed appropriate to reduce the overall number of epochs to avoid the risk of overfitting.

One of the key issues encountered in developing the phrase extraction part of this study is that the dataset is heavily slanted toward imbalanced classes and also a token-level annotated dataset. It was confirmed that the use of subword tokens caused many words to become tokenized into individual subwords (tokens), which caused many words to have multiple tokens and that it was necessary to utilize exact label aligning strategies in order to avoid any noise in terms of the labels. Additionally, this study revealed that the validation results in this study exhibited extreme variance and were not consistently stable across all folds until the appropriate fold configurations and the number of epochs were appropriately defined and aligned.

In contrast, the sentiment classification portion of this study was considerably less difficult. Because the extracted phrases were short and well defined, the training resulted in rapid convergence and the experimentation produced a reliable level of convergence across all architectures.Previous research on token classification has shown that misclassification often occurs when labels share subtle contextual similarities or overlapping boundaries [9].

Chapter 8
CONCLUSION

This project successfully developed a two-stage NLP framework for fine-grained opinion mining by combining phrase extraction with sentiment classification. The first stage focused on identifying meaningful opinion-bearing phrases using transformer-based NER models such as RoBERTa, XLM-RoBERTa, BART, and BERT. Through 6-fold cross-validation and careful tuning to address overfitting, the system achieved stable performance with an overall accuracy of 0.82% and macro and weighted F1-scores of 0.72 and 0.82 respectively. The second stage applied the extracted phrases to sentiment classification models, which performed exceptionally well, achieving around 99% accuracy and a 99% F1-score, demonstrating the effectiveness of separating phrase extraction from sentiment prediction.

Overall, the proposed two-stage architecture proved to be more robust and interpretable than conventional sentence-level sentiment analysis. By isolating the exact opinion-bearing segments before classifying their polarity, the system provides more detailed insights and improves predictive reliability. The results confirm that transformer-based models, when combined with structured preprocessing and cross-validated training, form a powerful solution for fine-grained sentiment understanding in user-generated text.The results of this study reinforce the idea that transformer-based token classification and sentiment models are now mature enough to be deployed in real-world opinion mining systems.

8.1 SCOPE OF FURTHER WORK

Although the proposed pipeline achieves strong results, several directions for future improvement remain:

1. **Handling Token-Level Annotation Challenges:** critical to the effectiveness of the phrase extraction stage. When handled incorrectly, the creation of subword prefixes can introduce noise and misalignment between the token and label. Future work should include extending the implementation of encoder-decoder or span-based natural language processing techniques to the processing of spans instead of just individual tokens.

2. **Expanding and Refining the Dataset:** The existing dataset is relatively limited and manually annotated, which resulted in some label skewing potentially affecting early training consistency, therefore creating a

larger dataset with broader and more realistic reviews would aid in enhancing generalization. Methods for Semi-supervised learning can also be utilized to increase the training corpus with little manual input.

In summary, while the developed system achieves strong quantitative performance and successfully validates the two-stage approach, several technical enhancements can further improve robustness, scalability, and real-world applicability.

REFERENCES

[1] **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.**, Attention Is All You Need, *NeurIPS*, 2017.

[2] **Devlin, J., Chang, M., Lee, K., & Toutanova, K.**, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *NAACL-HLT*, 2019.

[3] **Mikolov, T., Chen, K., Corrado, G., & Dean, J.**, Efficient Estimation of Word Representations in Vector Space, *arXiv:1301.3781*, 2013.

[4] **Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., et al.**, Unsupervised Cross-lingual Representation Learning at Scale, *ACL*, 2020.

[5] **Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S.**, SemEval Aspect-Based Sentiment Analysis, *SemEval*, 2014.

[6] **Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., et al.**, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, *ACL*, 2020.

[7] **Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J.**, Deep Learning Based Text Classification: A Comprehensive Review, *ACM Computing Surveys*, 2021.

[8] **Kohavi, R.**, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, *IJCAI*, 1995.

[9] **Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C.**, Neural Architectures for Named Entity Recognition, *NAACL*, 2016.

[10] **Zhang, Z., & Liu, B.**, Aspect-Based Sentiment Analysis: A Survey, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2022.

[11] **Floridi, L., et al.**, Natural Language Processing in Artificial Intelligence: Trends and Opportunities, *Minds & Machines*, 2020.

[12] **Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., et al.**, RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv:1907.11692*, 2019.

[13] **Cambria, E.**, Affective Computing and Sentiment Analysis, *IEEE Intelligent Systems*, 2016.

[14] **Sun, C., Huang, L., & Qiu, X.**, Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentences, *NAACL*, 2019.

[15] **Zhang, Y., et al.**, Contrastive Learning for Sentiment Analysis, *ACL*, 2021.

[16] **Yu, J., Jiang, Z., Yang, Q., & Yang, J.**, Hierarchical Attention Mechanisms for Document Classification, *NAACL*, 2019.

[17] **Ribeiro, M. T., Singh, S., & Guestrin, C.**, ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier, *KDD*, 2016.

[18] **Sundararajan, M., Taly, A., & Yan, Q.**, Axiomatic Attribution for Deep Networks, *ICML*, 2017.

[19] **Chefer, H., Gur, S., & Wolf, L.**, Transformer Interpretability Beyond Attention Visualization, *CVPR*, 2021.

[20] **Jin, X., et al.**, Is Attention Explanation? Analyzing Syntactic and Semantic Concepts Learned by BERT, *ACL*, 2020.