**The Impact of Code Review Sentiment on Pull Request Success: A Comparative Analysis of Selected Pull Requests on GitHub**

[1. Introduction 4](#_Toc190951780)

[1.1. Background and Motivation 4](#_Toc190951781)

[1.2. Research Objectives 4](#_Toc190951782)

[1.3. Research Questions 4](#_Toc190951783)

[1.4. Scope and Limitations 4](#_Toc190951784)

[1.5. Research Contributions 4](#_Toc190951785)

[1.6. Thesis Structure 4](#_Toc190951786)

[2. Related Work 4](#_Toc190951787)

[2.1. Overview of Existing Studies 4](#_Toc190951788)

[2.1.1. Prior Research on Code Review Sentiment 4](#_Toc190951789)

[2.1.2. Research Gaps & Limitations 4](#_Toc190951790)

[2.2. Terminology and Definitions 4](#_Toc190951791)

[2.2.1. Sentiment analysis 4](#_Toc190951792)

[ Definition 4](#_Toc190951793)

[ Application 4](#_Toc190951794)

[ Challenges 4](#_Toc190951795)

[2.2.2. Deep Learning Models 4](#_Toc190951796)

[ Transformer Models 4](#_Toc190951797)

[ DistilBERT, DeBERTa, CodeBERT 4](#_Toc190951798)

[ Fine-tuning 4](#_Toc190951799)

[ Model Performance 4](#_Toc190951800)

[2.2.3. Code Review and PR Success 4](#_Toc190951801)

[ Definition and Importance 4](#_Toc190951802)

[ PR Success Metrics 4](#_Toc190951803)

[ Factors Affecting PR Acceptance 4](#_Toc190951804)

[2.2.4. Statistical Methods 4](#_Toc190951805)

[ Chi-Square Test 4](#_Toc190951806)

[ Cramér’s V 4](#_Toc190951807)

[ Effect Size 4](#_Toc190951808)

[3. Methodology 4](#_Toc190951809)

[3.1. Research Approach 4](#_Toc190951810)

[3.2. Model Selection 5](#_Toc190951811)

[3.2.1. Training Set 6](#_Toc190951812)

[3.2.2. Preprocessing data 6](#_Toc190951813)

[3.2.3. Fine-Tuning 6](#_Toc190951814)

[3.2.4. Model Evaluation 6](#_Toc190951815)

[3.3. Data Collection 6](#_Toc190951816)

[3.3.1. Sample Size Calculation 7](#_Toc190951817)

[3.3.2. Data Collection criteria 7](#_Toc190951818)

[3.3.3. Data Cleaning 7](#_Toc190951819)

[3.3.4. Collected Information 7](#_Toc190951820)

[3.4. Sentiment classification 8](#_Toc190951821)

[3.4.1. Comments and Reviews Sentiment Classification 8](#_Toc190951822)

[3.4.2. PRs Sentiment Measurement 8](#_Toc190951823)

[3.4.3. PR success Metrics 9](#_Toc190951824)

[3.5. Hypothesis Testing 9](#_Toc190951825)

[3.6. Validity and Reliability 9](#_Toc190951826)

[4. Results and Findings 9](#_Toc190951827)

[4.1. Labelled Data Preparation 9](#_Toc190951828)

[4.1.1. Dataset Balancing 9](#_Toc190951829)

[4.1.2. Data Cleaning and Preprocessing 9](#_Toc190951830)

[4.1.3. Data Splitting and Sampling Strategy 9](#_Toc190951831)

[4.2. Model Evaluation 9](#_Toc190951832)

[4.2.1. Model Selection 9](#_Toc190951833)

[4.2.2. Training and Fine-Tuning 9](#_Toc190951834)

[4.2.3. Baseline Comparison 9](#_Toc190951835)

[4.2.4. Performance Metrics and Error Analysis 9](#_Toc190951836)

[4.3. PR Data Collection 10](#_Toc190951837)

[4.3.1. Retrieving PRs, Comments, and Reviews 10](#_Toc190951838)

[4.3.2. Sample Size Justification and Power Analysis 10](#_Toc190951839)

[4.3.3. Data Reliability and Validity 10](#_Toc190951840)

[4.3.4. Data Processing and Preparation 10](#_Toc190951841)

[4.4. Sentiment Analysis on PR Data 10](#_Toc190951842)

[4.4.1. Sentiment Classification of PR Comments and Reviews 10](#_Toc190951843)

[4.4.2. Aggregated PR Sentiment Scores 10](#_Toc190951844)

[4.4.3. PR Success Measurement 10](#_Toc190951845)

[4.5. Statistical Analysis and Hypothesis Testing 10](#_Toc190951846)

[4.5.1. Data Preparation for Statistical Testing 10](#_Toc190951847)

[4.5.2. Findings for Research Question #1 10](#_Toc190951848)

[4.5.3. Findings for Research Question #2 10](#_Toc190951849)

[4.5.4. Hypothesis Testing Results 10](#_Toc190951850)

[4.5.5. Chi-Square and Effect Size Results 10](#_Toc190951851)

[5. Discussion 10](#_Toc190951852)

[5.1. Interpretation of Results 10](#_Toc190951853)

[5.1.1. Model Performance Analysis 10](#_Toc190951854)

[5.1.2. Impact of Sentiment on PR Success 10](#_Toc190951855)

[5.2. Comparison with Prior Studies 10](#_Toc190951856)

[5.3. Discussion of Limitations 10](#_Toc190951857)

[5.3.1. Limitations in Sentiment Classification Models 10](#_Toc190951858)

[5.3.2. Uncontrolled External Factors Affecting PR Success 10](#_Toc190951859)

[5.4. Threats to Validity 10](#_Toc190951860)

[6. Conclusion 10](#_Toc190951861)

[6.1. Summary 10](#_Toc190951862)

[6.2. Recommendations for Future Research 10](#_Toc190951863)

# Introduction

## Background and Motivation

## Research Objectives

## Research Questions

## Scope and Limitations

## Research Contributions

## Thesis Structure

# Related Work

## Overview of Existing Studies

### Prior Research on Code Review Sentiment

### Research Gaps & Limitations

## Terminology and Definitions

### Sentiment analysis

#### Definition

#### Application

#### Challenges

### Deep Learning Models

#### Transformer Models

#### DistilBERT, DeBERTa, CodeBERT

#### Fine-tuning

#### Model Performance

### Code Review and PR Success

#### Definition and Importance

#### PR Success Metrics

#### Factors Affecting PR Acceptance

### Statistical Methods

#### Chi-Square Test

#### Cramér’s V

#### Effect Size

# Methodology

## Research Approach

This study adopted a hybrid research approach, including:

* An experimental approach to fine-tune, evaluate, and compare deep learning models for sentiment analysis and select the model with the highest performance in classifying the Software Engineering texts.
* A quantitative approach for data collection from GitHub open source projects, sentiment classification of PRs, and statistical hypothesis testing.

Figure 1 illustrates the methodology workflow employed in this study.

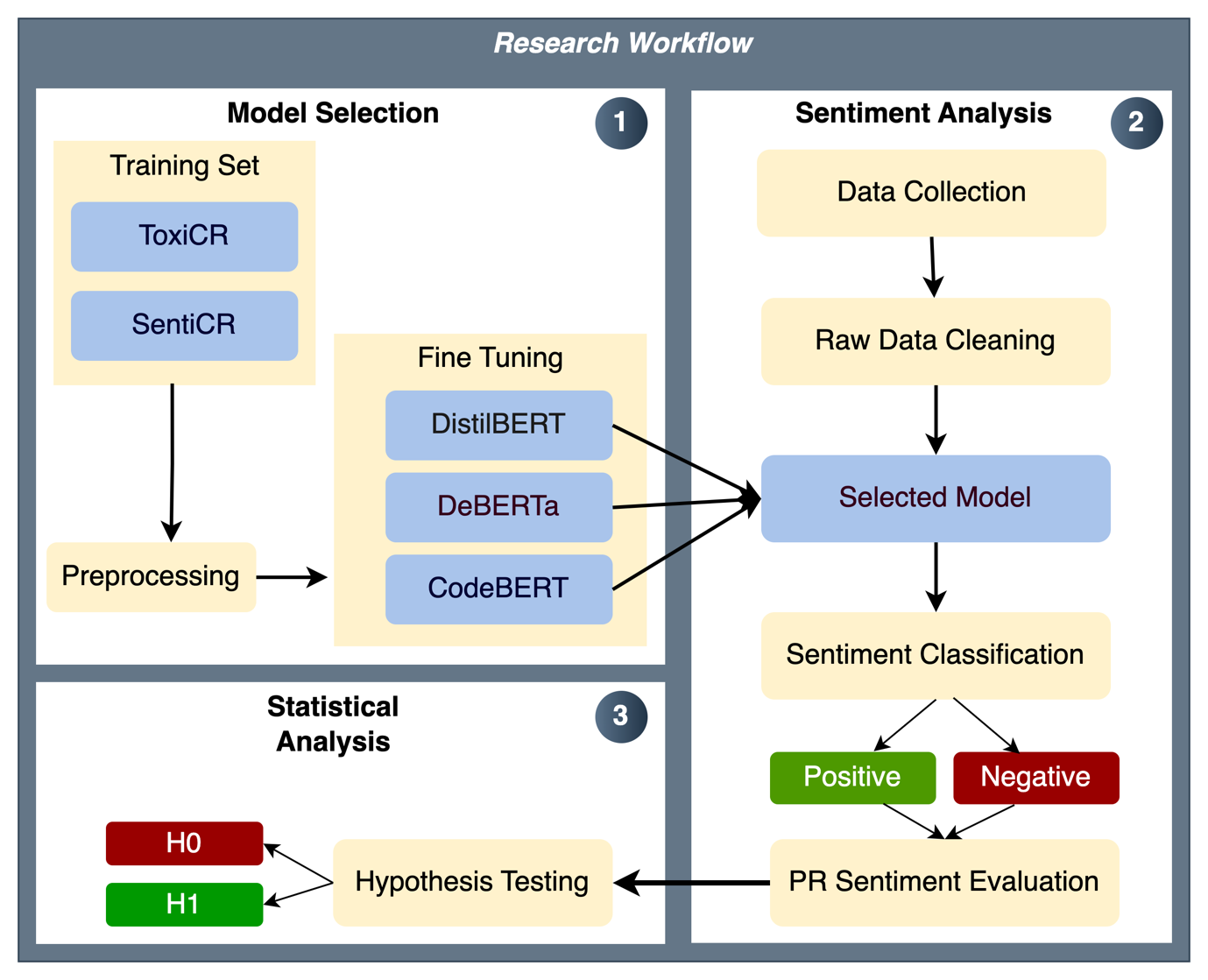


Figure 1. Research Workflow

This chapter explains the steps undertaken, including data collection, model selection, sentiment classification, PR sentiment measurement, and comparative analysis.

## Model Selection

Three transformer models were selected, including DistilBERT, DeBERTa, and CodeBERT. DistilBERT was chosen for its efficiency, DeBERTa for its strong NLP performance, and CodeBERT for its specialized understanding of SE-related text.

### Training Set

The transformers were retrained using publicly available software engineering-specific datasets.

* **Oracle dataset** (SentiCR): Contains 1,600 GitHub reviews (Ahmed et al., 2017).
* **ToxiCR dataset**: Contains 19,651 GitHub code review comments (Sarker et al., 2023).

A training, validation, and test split of 75:15:15 was applied, with stratified sampling to maintain class balance.

### Preprocessing data

Preprocessing was performed on both datasets to remove noise and prepare them as input for deep learning models. The steps include:

* **Text Cleaning**: Removing URLs and handling special characters and emojis.
* **Tokenization and Vectorization**: Using the default tokenization methods from transformer models.

The cleaned and vectorized text was fed into the transformer models for training and classification.

### Fine-Tuning

All the models underwent a fine-tuning process, where hyperparameters were adjusted to achieve optimal performance.

### Model Evaluation

The models were evaluated using standard classification metrics:

* Accuracy: Percentage of correctly classified instances.
* Precision: Proportion of positive classifications that are positive.
* Recall: Ability of the model to identify all relevant positive instances.
* F1-Score: Harmonic mean of precision and recall.

The best-performing classifier was used to analyze the data from GitHub and conduct a comparative analysis.

## Data Collection

The data for this study was collected from GitHub open source repositories using the GitHub REST APIs and a Python script to automate the process, ensuring reproducibility.

### Sample Size Calculation

The sample size was calculated using Cohen’s d to ensure sufficient statistical power. Based on sentiment classification, this determines the minimum number of PRs needed to detect meaningful differences in PR success.

However, in this study, the sample size exceeded the required minimum calculated using Cohen’s d, and 1,000 PRs from 10 repositories were collected.

### Data Collection criteria

PRs from different open-source projects were selected based on the following conditions:

* Only closed PRs were included to ensure measurable success (merged/not merged).
* The 100 most-commented PRs from each repository were selected to capture PRs with high levels of communication.
* Comments and reviews for each PR were collected until the PR’s closure.

The focus on most-commented PRs was intentional, as these discussions tend to contain more heated exchanges and, therefore, provide a more affluent sample of negative sentiments, typically underrepresented in general PR discussions. This sampling strategy ensured sufficient data for meaningful comparative analysis of positive and negative sentiment impacts.

### Data Cleaning

To ensure relevant data, the following preprocessing steps were applied:

* **Removal of bot-generated comments**: PR comments that are generated by bots do not represent human-to-human developer communication. Therefore, these were identified and removed through manual repository inspection and identifying bot names.
* **Exclusion of empty reviews**: Reviews that contain no textual feedback were discarded, as they did not contribute to sentiment analysis.

### Collected Information

The extracted data contained the following information:

Table 2. Metadata of Pull Requests

|  |  |
| --- | --- |
| **Column** | **Description** |
| id | Unique identifier of the PR |
| number | PR number in the repository |
| title | Title of the PR |
| user | Username of the PR author |
| created\_at | Timestamp when PR was created |
| merged\_at | Timestamp when PR was merged (if applicable) |
| comments | Number of comments on the PR |
| state | State of the PR (closed, merged) |
| closed\_at | Timestamp when PR was closed, whether it was merged or not. |

Table 3. Metadata of PR Comments

|  |  |
| --- | --- |
| **Column** | **Description** |
| id | Unique identifier of the comment |
| pr\_number | PR number the comment belongs to |
| user | Username of the commenter |
| created\_at | Timestamp when the comment was posted |
| body | Content of the comment |

Table 4. Metadata of PR Reviews

|  |  |
| --- | --- |
| **Column** | **Description** |
| id | Unique identifier of the review |
| pr\_number | PR number the review belongs to |
| user | Username of the reviewer |
| submitted\_at | Timestamp when the review was submitted |
| state | State of the review (Approved, Changes\_Requested, Commented) |
| body | Content of the Review |

## Sentiment classification

### Comments and Reviews Sentiment Classification

The trained models will classify each comment and review as **positive** or **negative.**

### PRs Sentiment Measurement

PR sentiment was determined using a majority voting mechanism, where the sentiment of individual comments and reviews was aggregated. If the sentiment classification resulted in equal positive and negative comments, the PR was labeled as neutral. Since neutral sentiment is not the focus of this study, these PRs were excluded from the analysis.

### PR success Metrics

The **merge rate** (whether a PR was merged or closed) was a key metric to determine PR success. This was calculated for each PR after sentiment measurement.

## Hypothesis Testing

This study examined whether sentiment influences PR success by testing the following hypotheses:

* **Null hypothesis (H₀):** Sentiment in PR comments does not influence the PR success.
* **Alternative hypothesis (H₁):** PRs with positive sentiment have a higher merge rate.

To evaluate this, a **Chi-square test of independence** was conducted to determine whether sentiment and PR success (merged/closed) are statistically dependent. The null hypothesis is rejected if the test yields a **p-value < 0.05**, indicating a significant relationship. Additionally, **Cramér’s V** was used to measure the strength of the association.

## Validity and Reliability

We acknowledge the **selection bias** in collecting only the most-commented PRs. As a result, the collected data may not be fully generalized to all developer communications on GitHub. However, this study ensures:

* **Internal Validity:** Only closed PRs were used, ensuring measurable success.
* **External Validity:** The inclusion of 10 diverse repositories enhanced the generalizability.
* **Reliability:** Automated data collection ensured consistency and reproducibility.

# Results and Findings

## Labelled Data Preparation

### Dataset Balancing

### Data Cleaning and Preprocessing

### Data Splitting and Sampling Strategy

## Model Evaluation

### Model Selection

### Training and Fine-Tuning

### Baseline Comparison

### Performance Metrics and Error Analysis

* Model Evaluation Metrics (Accuracy, Precision, Recall, F1-score)
* **Error Analysis and Misclassification Review**

## PR Data Collection

### Retrieving PRs, Comments, and Reviews

### Sample Size Justification and Power Analysis

### Data Reliability and Validity

### Data Processing and Preparation

## Sentiment Analysis on PR Data

### Sentiment Classification of PR Comments and Reviews

### Aggregated PR Sentiment Scores

### PR Success Measurement

## Statistical Analysis and Hypothesis Testing

### Data Preparation for Statistical Testing

### Findings for Research Question #1

### Findings for Research Question #2

### Hypothesis Testing Results

### Chi-Square and Effect Size Results

# Discussion

## Interpretation of Results

### Model Performance Analysis

### Impact of Sentiment on PR Success

## Comparison with Prior Studies

## Discussion of Limitations

### Limitations in Sentiment Classification Models

### Uncontrolled External Factors Affecting PR Success

## Threats to Validity

# Conclusion

## Summary

## Recommendations for Future Research