UNT
UNIVERSITY OF NORTH TEXAS

IARIA

# LIME-Aided Automated Usability Issue Detection from User Reviews: Leveraging LLMs for Enhanced User Experience Analysis


**Presented by**
Bassam Alsanousi
bassamalsanousi@my.unt.edu

**Research Authors:**
Bassam Alsanousi
Dr. Stephanie Ludi (Major professor)
Dr. Hyunsook Do (Co-Major Professor)

# About the Presenter

**Current Role:**

- Ph.D. Candidate at University of North Texas
- Specializing in Computer Science and Engineering

**Previous Experience:**

- IT Director at Technical and Vocational Training Corporation (TVTC)
- Lecturer at (TVTC)

**Research Interests:**

- Focused on Human-Computer Interaction (HCI), Mining Software Repositories, NLP, AI, and LLMs.

**Current Research:**

- Empowering the Analysis of Mobile Apps User Feedback Leveraging LLMs

# Outlines

1. Research Overview

2. Methodology

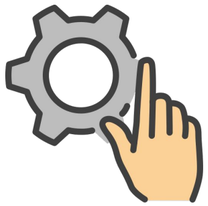
3. Results

4. Discussion and threats to validity

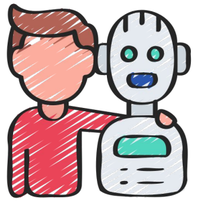
5. Conclusion

# Research Overview

# Problem Statement

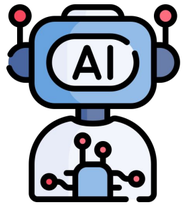


- Most of the existing studies have analyzed user reviews of mobile apps to pinpoint usability challenges through manual or semi-automated approaches

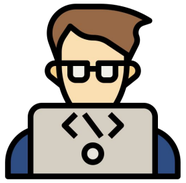


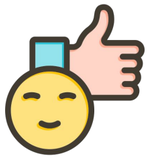
- These efforts underscore the growing need for more semantically-aware techniques, aiding developers in refining the quality of mobile apps

# Research Contributions




- We contribute by developing an approach that automatically detects usability issues from user reviews by leveraging LLMs



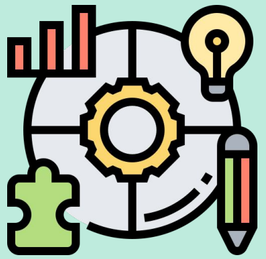


- This approach will provide developers with a more efficient method to identify usability concerns
- Aiming to improve the quality of mobile applications and enhance the overall user experience (UX)

# Research Questions

| | |
|---|---|
| **RQ1** | How effectively can LLMs semantically detect usability issues related to effectiveness, efficiency, and satisfaction from user reviews? |
| **RQ2** | Which LLMs have the most accurate results in classifying usability issues from user reviews? |
| **RQ3** | How do the classification from pre-trained models via API such as, (GPT-3.5 and GPT-4) by OpenAi and Llama 2 by Meta, compare to fine-tuned LLMs? |
| **RQ4** | How does applying explanation techniques such as local interpretable model-agnostic (LIME) enhance understanding model predictions for detecting usability issues? |

# Methodology



# Methodology - Usability Factors (ISO 9241-11)

| Usability Factors | Definition |
|---|---|
| Effectiveness | Assesses the users' ability to achieve their goals accurately and completely. Focuses on the extent to which users can accomplish their objectives. |
| Efficiency | Evaluates the level of effectiveness relative to the resources expended. Helps determine how efficiently users can attain their goals. |
| Satisfaction | Measures users' overall comfort and attitudes toward the product's usage. Reflects how users find the product's usage enjoyable and satisfactory. |

**Table 1:** Usability Factors

# Methodology - Examples

| Multi Classes | User Reviews Examples |
|---|---|
| Effectiveness, Efficiency, and Satisfaction | ”The new update is bad and the app is slow and sometimes gives errors” |
| Satisfaction | ”The worst banking app in the world.” |
| Efficiency | ”The application is slow and takes a long time to open and navigation between menus is slow” |
| Effectiveness | ”The application needs new maintenance and a new update. The amount does not appear in the account.” |
| Satisfaction and Effectiveness | ”The app keeps crashing. It’s very frustrating.” |

**Table 2:** Examples of Multi-Class Classification and Corresponding User Reviews as Usability Issues

# Methodology - Dataset

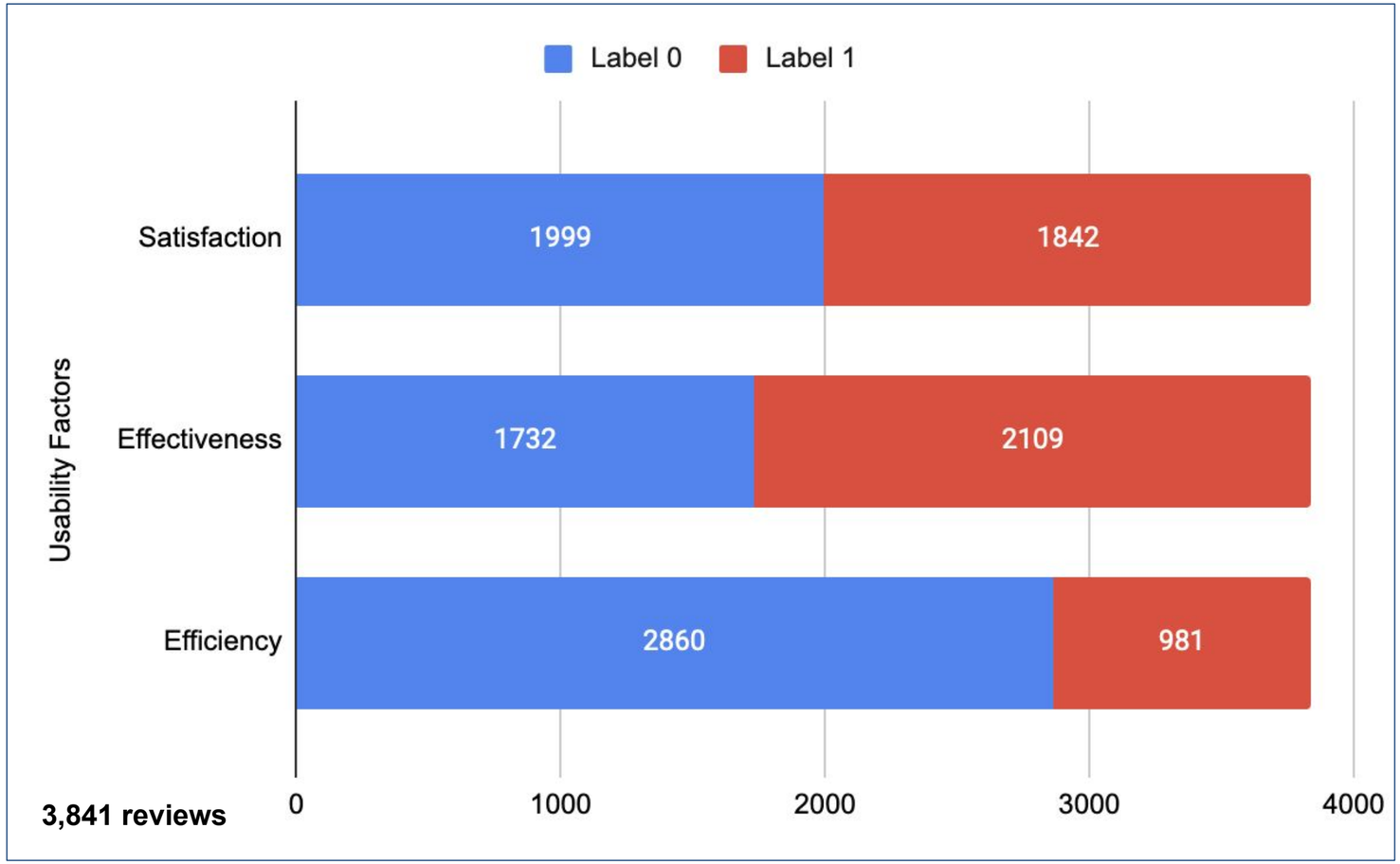



**Figure 1:** Dataset

# The LLMs employed in our research

| Model | Architecture | Parameters | Layers |
|---|---|---|---|
| BERT | bert-base-cased | 110M | 12 |
| RoBERTa | roberta-base | 125M | 12 |
| BART | bart-base | 140M | 6 |
| TinyBERT | General 4L 312D | 14M | 4 |
| XLNet | xlnet-base-cased | 110M | 12 |
| DistilBERT | distilbert-base-cased | 65M | 6 |
| GPT2 | gpt2 | 117M | 12 |

**Table 3:** Details About the Fine-tuned Large Language Models Employed in Our Research

GPT by OpenAI

Llama 2 by Meta

# Approach

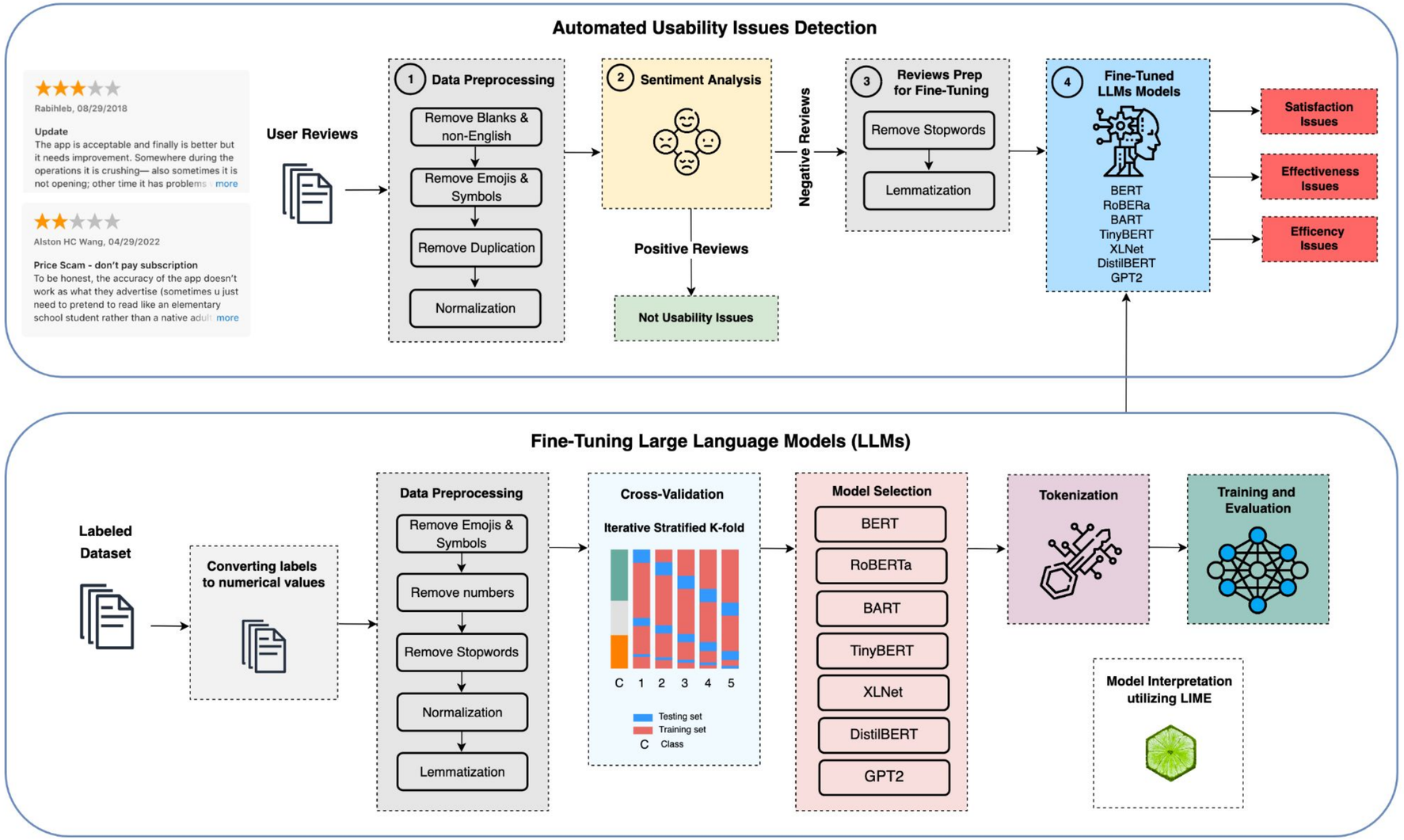



**Figure 2:** Proposed Approach

# Evaluation Metrics

Accuracy

Precision

Recall

F1

# Results

# Study Results

**RQ1:** LLMs can effectively and semantically detect usability issues related to effectiveness, efficiency, and satisfaction from user reviews.

| Model | Accuracy | Precision | Recall | F1 | Training Time (s) |
|---|---|---|---|---|---|
| BERT | 0.95 | 0.96 | 0.94 | 0.95 | 1645 |
| RoBERTa | **0.96** | 0.96 | 0.97 | 0.96 | 1336 |
| BART | 0.95 | 0.94 | 0.95 | 0.95 | 1619 |
| TinyBERT | 0.90 | 0.89 | 0.90 | 0.89 | **173** |
| XLNet | **0.96** | 0.95 | 0.96 | 0.95 | 1616 |
| DistilBERT | **0.96** | 0.96 | 0.96 | 0.96 | **806** |
| GPT2 | 0.93 | 0.92 | 0.93 | 0.92 | 1526 |

**Table 5:** The Results of each LLM Model

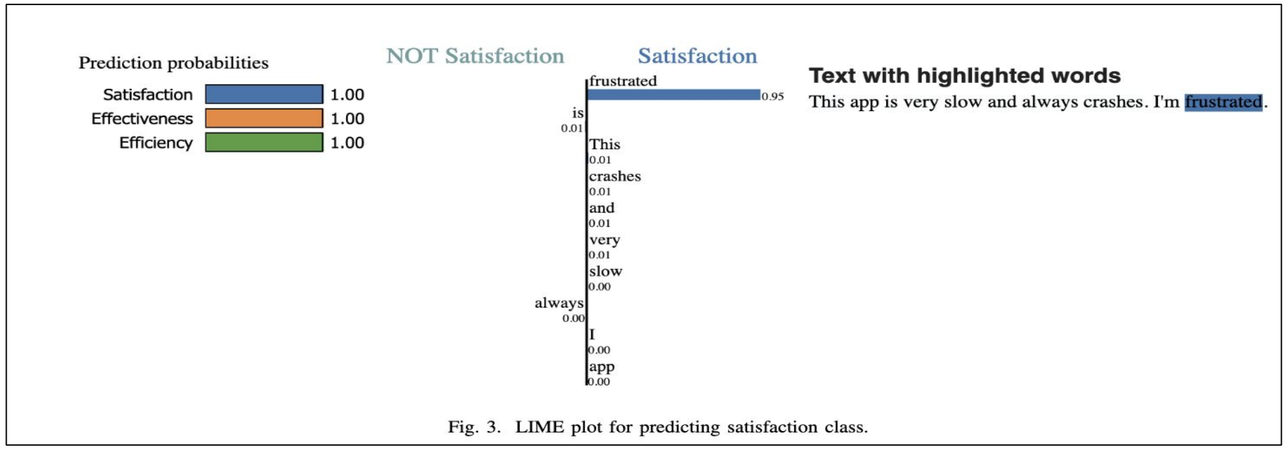



**Figure 10:** LIME Plot for Predicting Satisfaction Class

# Study Results

**RQ2:** RoBERTa, XLNet and **DisilBERT** have the most accurate results in classifying usability issues from user reviews.

| Model | Accuracy | Precision | Recall | F1 | Training Time (s) |
|---|---|---|---|---|---|
| BERT | 0.95 | 0.96 | 0.94 | 0.95 | 1645 |
| RoBERTa | **0.96** | 0.96 | 0.97 | 0.96 | 1336 |
| BART | 0.95 | 0.94 | 0.95 | 0.95 | 1619 |
| TinyBERT | 0.90 | 0.89 | 0.90 | 0.89 | **173** |
| XLNet | **0.96** | 0.95 | 0.96 | 0.95 | 1616 |
| DistilBERT | **0.96** | 0.96 | 0.96 | 0.96 | **806** |
| GPT2 | 0.93 | 0.92 | 0.93 | 0.92 | 1526 |

**Table 5:** The Results of each LLM Model

# Study Results

**RQ2:** RoBERTa, XLNet and **DisilBERT** have the most accurate results in classifying usability issues from user reviews.

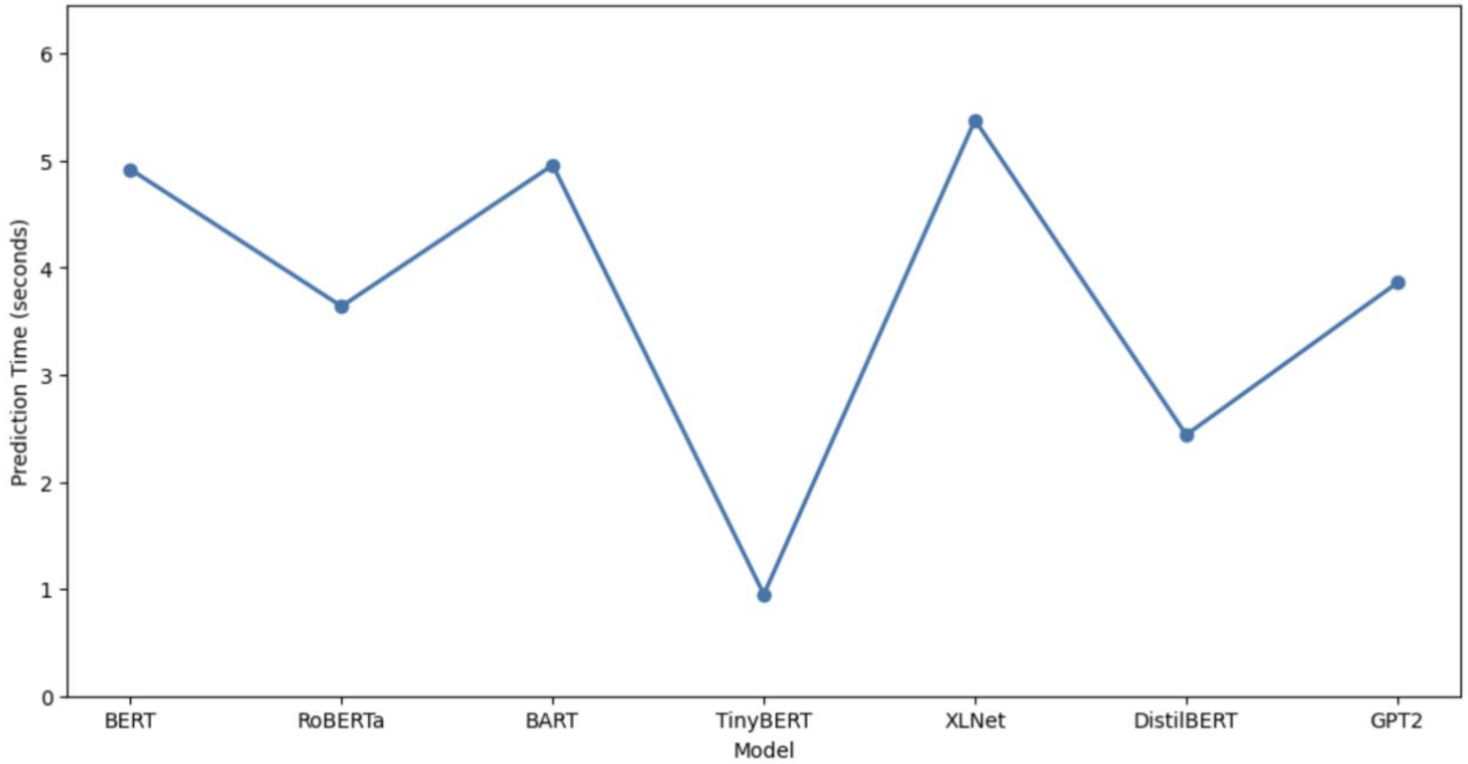



**Figure 11:** Prediction Times of Each Model

# Study Results

**RQ2:** RoBERTa, XLNet and **DisilBERT** have the most accurate results in classifying usability issues from user reviews.

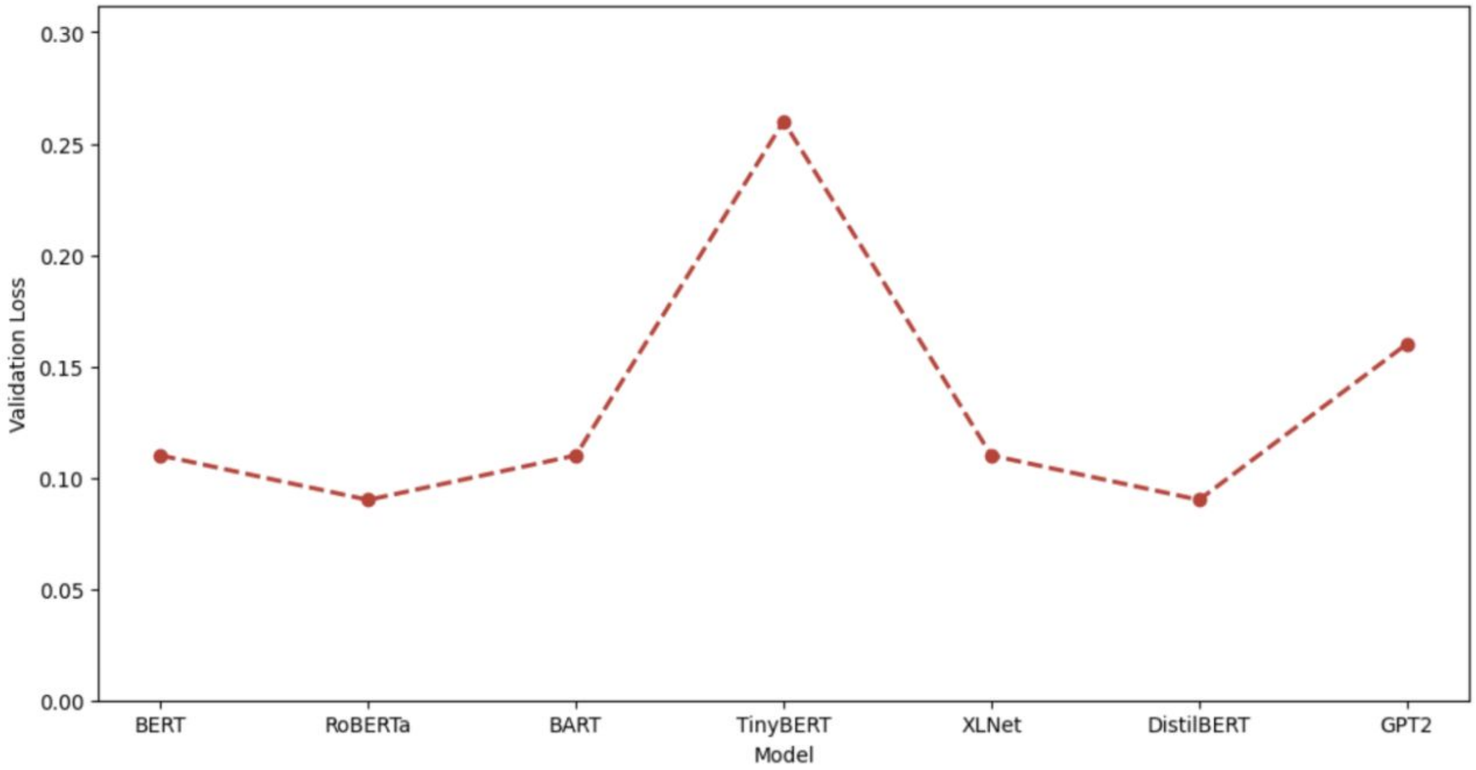



**Figure 12:** Validation Loss of Each Model

# Study Results

**RQ3:** Fine-tuned LLMs outperformed the pre-trained models via API such as, (GPT-3.5 and GPT-4) by OpenAI and Llama 2 by Meta.

| Model | Accuracy | Precision | Recall | F1 | Training Time (s) |
|---|---|---|---|---|---|
| BERT | 0.95 | 0.96 | 0.94 | 0.95 | 1645 |
| RoBERTa | **0.96** | 0.96 | 0.97 | 0.96 | 1336 |
| BART | 0.95 | 0.94 | 0.95 | 0.95 | 1619 |
| TinyBERT | 0.90 | 0.89 | 0.90 | 0.89 | **173** |
| XLNet | **0.96** | 0.95 | 0.96 | 0.95 | 1616 |
| DistilBERT | **0.96** | 0.96 | 0.96 | 0.96 | **806** |
| GPT2 | 0.93 | 0.92 | 0.93 | 0.92 | 1526 |
| Llama 2 - Zero-shot | 0.41 | 0.86 | 0.71 | 0.74 | - |
| Llama 2 - Few-Shot | 0.73 | 0.88 | 0.97 | 0.90 | - |
| GPT-3.5 | 0.64 | 0.89 | 0.89 | 0.86 | - |
| GPT-4 | 0.74 | 0.88 | 0.97 | 0.91 | - |

**Table 6:** The Results of each LLM Model

# Study Results

**RQ4:** LIME helped enhance understanding of model predictions for detecting usability issues in all classes.

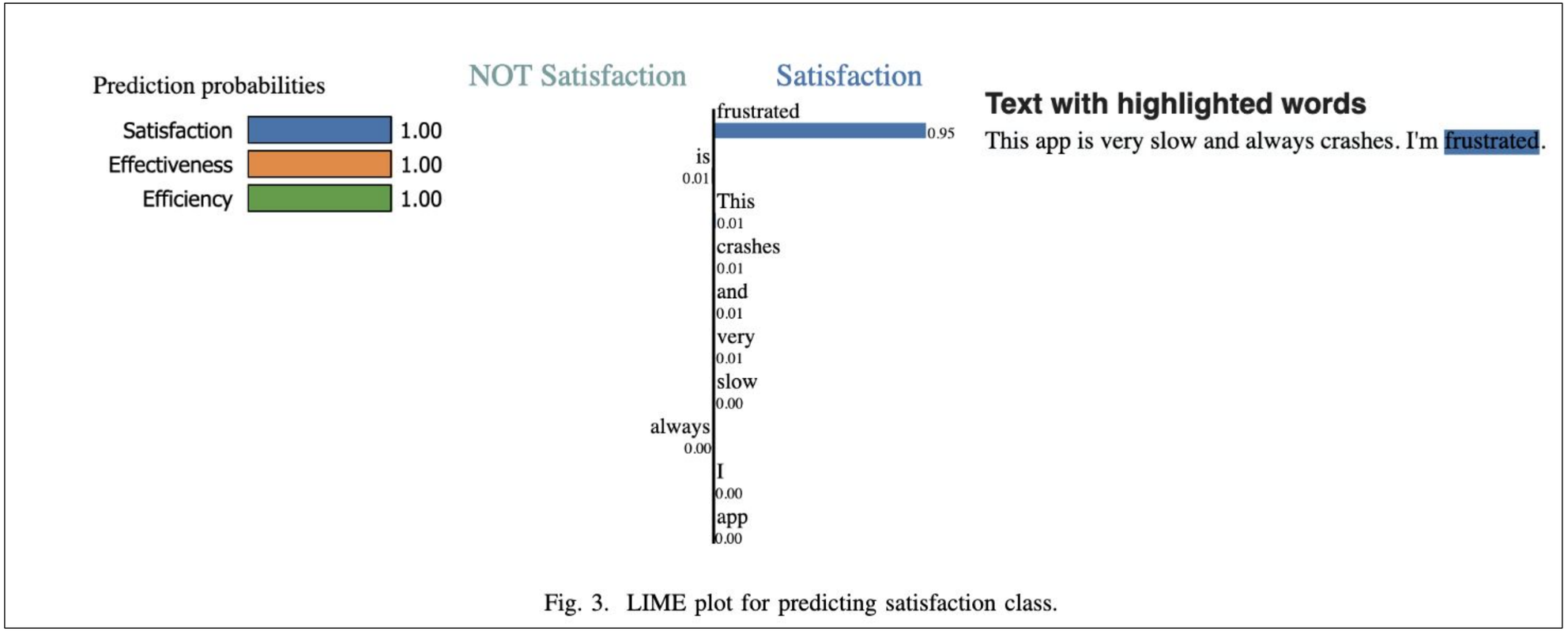



Fig. 3. LIME plot for predicting satisfaction class.

**Figure 10:** LIME Plot for Predicting Satisfaction Class

# Study Results

**RQ4:** LIME helped enhance understanding of model predictions for detecting usability issues in all classes.

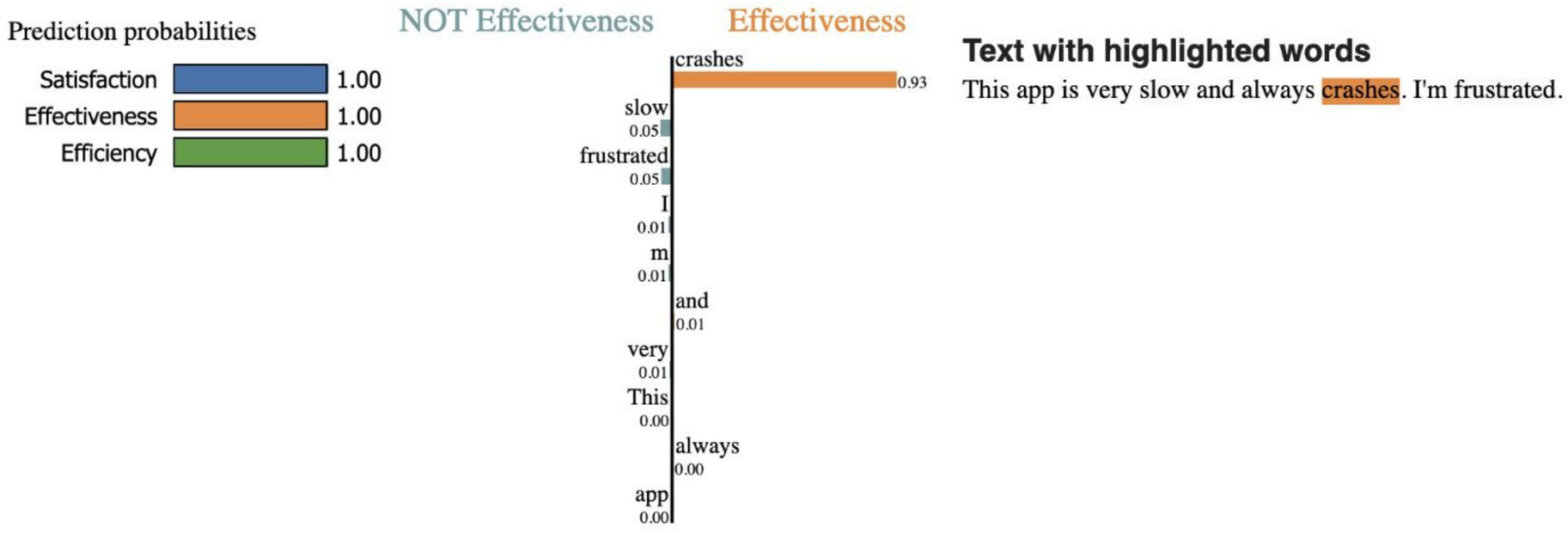



Fig. 4. LIME plot for predicting effectiveness Class.

**Figure 13:** LIME Plot for Predicting Effectiveness Class

# Study Results

**RQ4:** LIME helped enhance understanding of model predictions for detecting usability issues in all classes.

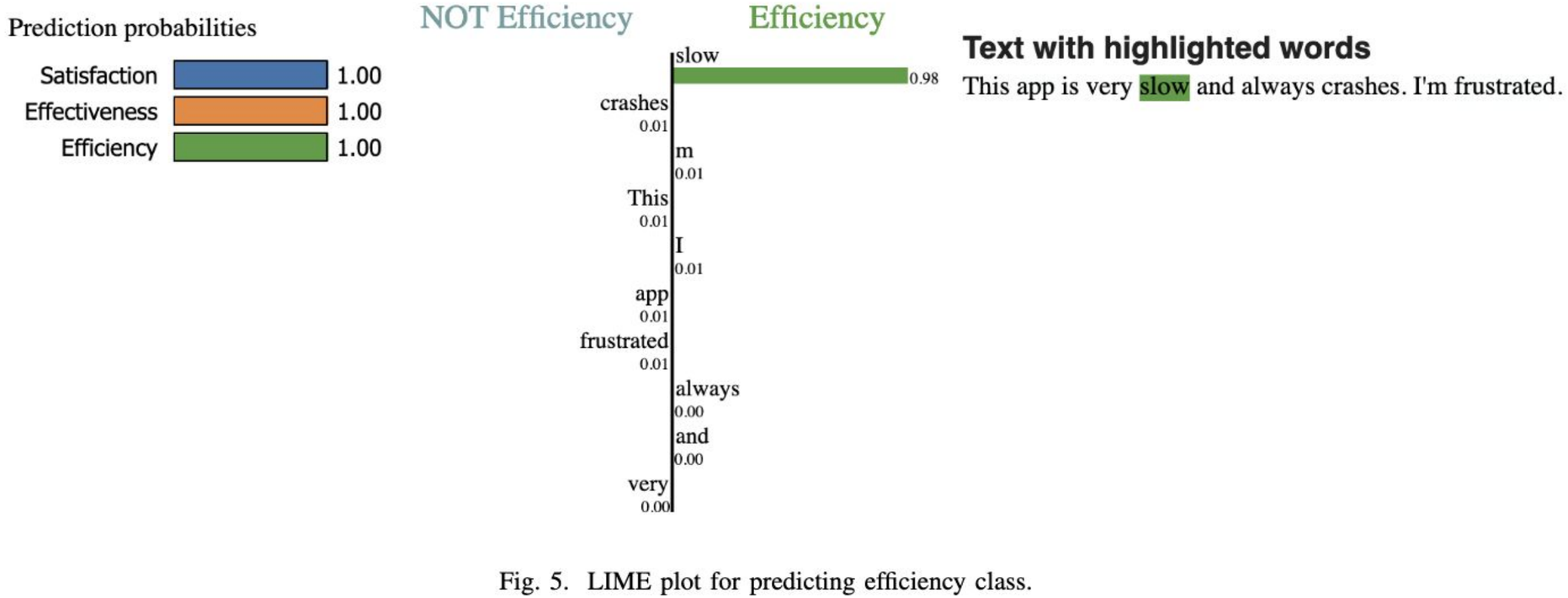



Fig. 5. LIME plot for predicting efficiency class.

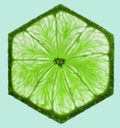
**Figure 14:** LIME Plot for Predicting Efficiency Class

# Discussion and threats to validity

# Discussion

Accuracy of Models

Predictive Reliability



Enhancement of Model Interpretability

Performance of Pre-trained Models

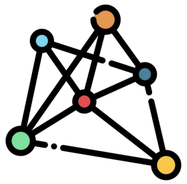
Model Training Efficiency

Real-world Applications of These Models

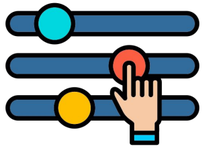
# Threats to validity

## Internal Validity

Model Overfitting

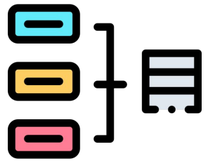


Parameter Tuning



## External Validity

Dataset Specificity



Language Bias

# Conclusion

# Conclusion and Future Work

- Demonstrated the significant role and effectiveness of LLMs in analyzing mobile apps usability to improve the UX
- Found that fine-tuned models, specifically RoBERTa, XLNet, and DistilBERT, outperformed others, including pre-trained models like GPT-3.5, GPT-4, and Llama 2, in detecting usability issues from user reviews
- Applied LIME for model interpretability, enhancing transparency and trustworthiness of fine-tuned models
- Future research will focus on applying these findings in a specific domain to deeply investigate usability and user experience (UX) of mobile app

# Thank you for listening