Understanding the Emotional Effects on GitHub Issue Lifecycles

İbrahim Enes Genişyürek, Oğuzhan Taşçı

Supervisor: Dr. Tuğba Gürgen Erdoğan
Computer Engineering, Hacettepe University

Introduction

❖ GitHub is a widely used platform for version control and collaboration, where developers create, manage, and discuss issues. Understanding the emotional tone of issues and comments can provide insights into the development process, improve issue management, and enhance team dynamics. This study explores the impact of emotions on the lifecycle of GitHub issues using NRC Emotion Lexicons, which categorize emotions into various dimensions like joy, anger, sadness, and trust.

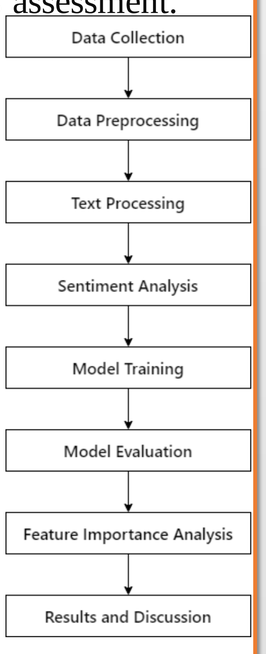
Specifications and Design Requirements

•**Specifications:**
- •**Data Collection:** Collecting GitHub issues and comments data and modelling them as events.
- •**Text Processing:** Tokenization, cleaning, and sentiment analysis.
- •**Modeling:** Using Random Forest and Gradient Boosting for predictive analysis.
- •**Evaluation:** Assessing model performance using metrics like MSE, R^2, MAE and MAPE.

•**Design Requirements:**
- •Handling large datasets efficiently.
- •Ensuring reproducibility and scalability of the analysis pipeline.

Solution Methodology

Software:
- **Data Preprocessing:** Cleaning and filtering data, removing bots.
- **Sentiment Analysis:** Using pre-trained sentiment models.
- **Feature Extraction:** TF-IDF for text data.
- **Model Training:** Hyperparameter tuning using Grid Search for Random Forest and Gradient Boosting models.
- **Evaluation:** Cross-validation and performance metrics assessment.



Theoretical Aspects:
- •Use of Natural Language Processing (NLP) techniques.
- •Machine learning algorithms for predictive analysis.

Practical Aspects:

•**Data Collection and Manipulation:**
- •**Pandas:** For data manipulation and analysis.
- •**NumPy:** For numerical operations and handling arrays.

•**Data Preprocessing:**
- •**re (Regular Expressions):** For cleaning and preprocessing
- •text data.

•**Text Processing and Feature Extraction:**
- •**NLTK (Natural Language Toolkit):** For natural language processing tasks such as tokenization and stopword removal.
- •**scikit-learn:** For feature extraction (e.g., TF-IDF), model training, evaluation, and machine learning algorithms.

•**Sentiment Analysis:**
- •**TextBlob:** For performing sentiment analysis.

•**Machine Learning:**
- •**scikit-learn:** For implementing machine learning models like Random Forest and Gradient Boosting, and for performing hyperparameter tuning with GridSearchCV.

•**Visualization:**
- •**Matplotlib:** For creating plots and visualizations.
- •**Seaborn:** For creating advanced statistical visualizations.

•**Evaluation:**
- •**scikit-learn metrics:** For calculating evaluation metrics like Mean Squared Error (MSE).

Application Areas

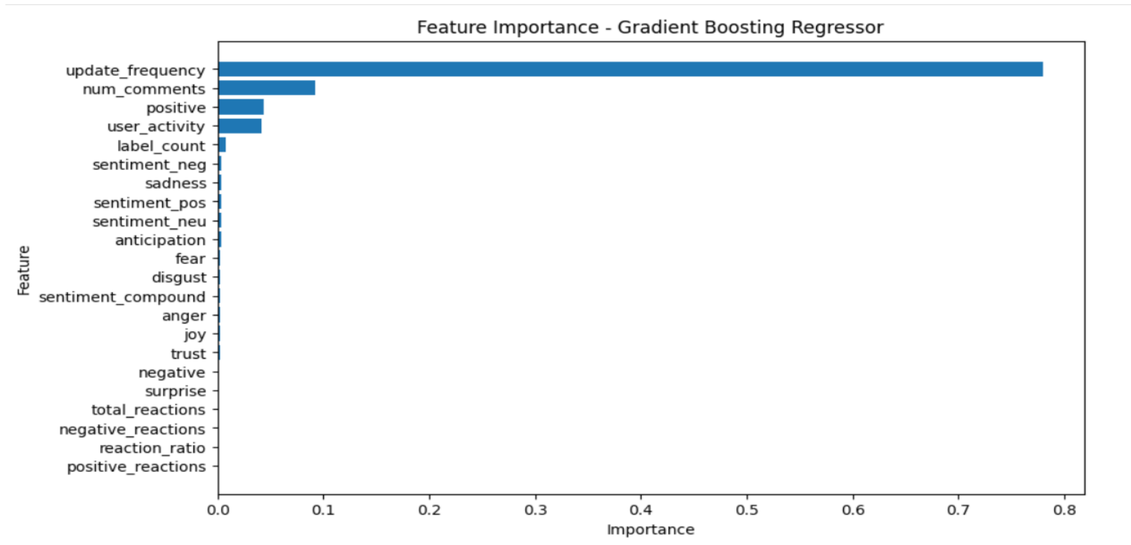
• **Software Development:**
Enhancing issue tracking systems.
Improving project management tools.
• **Research:**
Contributing to studies in software engineering and human-computer interaction.
• **Industry:**
Providing insights for tech companies to streamline their development processes.

Results and Discussion

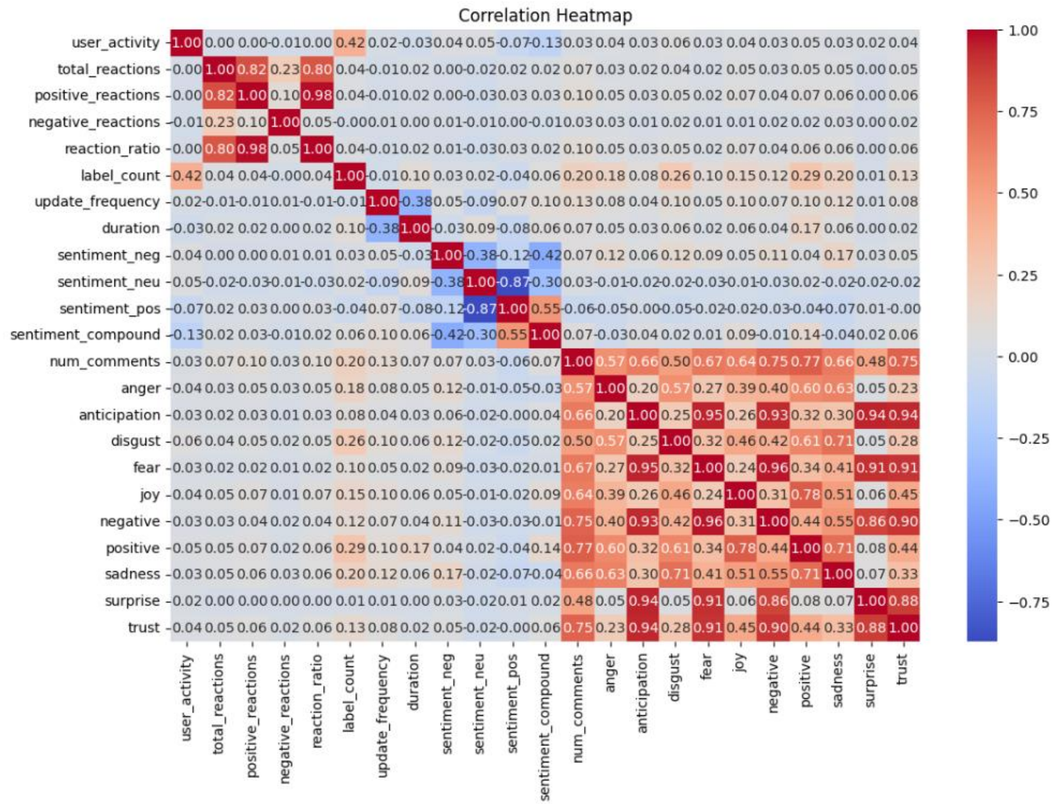
```
Cross-Validation Results:
fit_time                                  2.483132
score_time                                0.009802
test_neg_mean_squared_error              -0.091560
test_r2                                   0.893022
test_neg_mean_absolute_error             -0.138336
test_neg_mean_absolute_percentage_error  -0.468324
```

- Gradient Boosting Regressor – MSE: 0.09540023940988968, R^2: 0.9088054582564025, MAE: 0.14462891975400652,
  MAPE: 0.4329152814907027

The cross-validation and Additional Metrics demonstrates strong predictive performance.



The feature importance analysis highlights that update frequency and number of comments are the most significant predictors in predictive model on GitHub issue lifecycles, with positive sentiment also contributing but to a lesser extent, while individual sentiment metrics and specific emotions have minimal impact, suggesting a focus on interaction metrics for managing issue resolutions.



The correlation heatmap analysis reveals that issues receiving more reactions, whether positive or negative, generally have a higher total reaction count, with positive reactions significantly influencing the reaction ratio. Emotions such as anticipation correlate highly with positive sentiment, while anger correlates with negative sentiment, indicating that emotionally charged issues drive community engagement.

Acknowledgements

❖ This project was completed within the context of BBM479-480 Design Project graduation courses in Hacettepe University, Faculty of Engineering, Department of Computer Engineering.
❖ We thank Dr. Tuğba Gürgen Erdoğan for her invaluable contributions to our project.