Github Issue Label And Description Suggestion

Javed Habib
12140830
javedh@iitbhilai.ac.in

Mohd Adil
12141080
mohdadil@iitbhilai.ac.in

Advised by: Dr. Gagan Raj Gupta

ABSTRACT

When open-source projects receive contributions and bug reports, they often involve managing a significant number of issues. Labeling issues correctly is essential for efficient issue tracking and resolution. However, it can be a time-consuming task. The aim of this project is to develop a machine learning model that can automatically predict appropriate labels for GitHub issues based on their content.

KEYWORDS

NLP, Text Classification, BERT

Reference Format:
Javed Habib and Mohd Adil. 2023. Github Issue Label And Description Suggestion. In *IIT Bhilai   2023, India.* 1 page.

1 MOTIVATION

- Automatic Label Prediction: Develop a machine learning model that can predict relevant issue labels based on the issue's title and description.
- Better Issue Annotations: Model should be able to guide developers to write better description based on the label.

2 DATA SET AND PAPER

Our project uses a Kaggle-sourced dataset, comprising GitHub issues, including their titles, descriptions, and labels. Our project draws inspiration from BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art natural language processing model.

3 METHODOLOGY

Our methodology follows a structured approach, beginning with data cleaning to ensure the quality and consistency of our GitHub issue dataset, which includes titles, descriptions, and labels. We leverage BERT, a leading natural language processing model, for feature engineering, utilizing its pretrained embeddings to capture nuanced contextual information from the text data. Fine-tuning BERT on our dataset is a pivotal step, enabling the model to adapt specifically to the task of issue label prediction. Post-training, we evaluate the model using metrics such as precision, recall, F1-score, and accuracy, ensuring it performs optimally. Hyperparameter tuning to fine-tune the model's efficiency and effectiveness.
Uniqueness

- Implement bidirectional modal to capture better context of the issue
- Feeback mechanism for user to suggest better issue description and title based on label.

4 CONTRIBUTION

- **Adil** : Conducting extensive research on BERT, pretrained models, and the domain of Natural Language Processing (NLP). Participating in the integration and optimization of the BERT model, specifically tailored for our classification task. This encompassed fine-tuning the model, optimizing hyperparameters, and rigorous validation procedures to ensure the accuracy and efficacy of our model.
- **Javed** : Leveraging the BERT pre-trained model to enhance the model's NLP capabilities, conducting thorough data cleaning to ensure the dataset's quality, and finally creating the main model for label prediction. Additionally, participating in fine-tuning the model to align it with the specific task of issue label prediction, engage in validation processes to evaluate its performance, and contribute to the critical task of hyperparameter tuning to optimize the model's efficiency.

This report is submitted to IIT Bhilai under prof. Dr. Gagan Raj Gupta for course work CS550

, ,