?

Aliaksei Bialiauski
aliaksei.bialiauski@hey.com
?
Minsk, Belarus

Abstract

This paper is about something new.

Keywords: Machine Learning, Text Classification, Random-Forest, Transformers

1 Introduction

TBD.. Author [1]

2 Related Work

3 Research Method

The goal of this study is to understand whether GitHub repositories can be classified as sample or real. This leads to the following research questions:

RQ1 Do text transformers can predict classes based on text?

RQ2 Which technique performs better in task to classify GitHub repositories on real and sample?

First, we prepared a training dataset of 1,000 public GitHub repositories. It is important to have both: real projects and repositories with examples. We distributed number of repositories between real and samples 750 and 250 respectively. Sample repositories were queried like that: 84 repositories that contain `examples` in their name, 83 repositories named with `samples`, and 83 contain `guides` in their name. For each GitHub repository we collected the following features: 1) `description`: repository's description 2) `readme`: README.md file 3) `created_at`: date when repository was created 4) `last_commit`: latest commit date 5) `commits`: total amount of commits

Second, we label our dataset using numeric labels. The real repository labeled as 0, while sample one labeled as 1. To automate the process of labeling we utilize pattern-matching script. We run pattern matching for repository's full name e.g.: `yegor256/takes`, `apache/kafka`, and it's description. For instance, `leeowenowen/rxjava-examples` will match, while `objectionary/eo` won't.

Third, we prepare dataset to be presented to the model as input. For this purpose we preprocess data using techniques of tokenization and stopwords removal. After this step, we feed this dataset by learning both, machine learning model with Random-Forest algorithm, and deep learning text transformers as motivated in RQ2.

Finally, we collect and compare the results, produced by trained models.

All GitHub repositories we collected are public repositories with more than 20 stars.

4 Results

TBD..

5 Limitations

6 Discussion

7 Conclusion

8 Acknowledgements

References

[1] Test Author. 2024. Test Article. *Test Journal* (2024).