Informatics Institute of Technology
In Collaboration with
University of Westminster, UK

"KNOWLEDGE MINING CLASSIFIER"

Mining and Classifying Posts from Q&A Platform on Issues to Improve Maintainability

A Dissertation by
Mr. Shuhaib Ali

Supervised By
Mr. Guhanathan Poravi

Submitted in partial fulfilment of the requirements for the
BEng (Hons) Software Engineering Degree
Department of Computing

April 2019

Abstract

Designing and maintaining of APIs, languages, libraries and frameworks is a complex task due to the repeatedly changing requirements from its current users. Incomplete, poor documentation, poor performance, lack of backward compatibility and bugs can be considered as some of the foremost issues faced. It shows a huge potential towards minimizing the number of issues caused. Q&A related platforms has become one of the most popular places for discussing the above stated issues, since it creates communities that draw in experts at a professional level. The main challenges found in the current Q&A platforms is the availability of unstructured posts which in result makes it difficult for the designers to identify relevant issues. The current/ existing systems that were developed only considered few features on the learning state such as NLP and labelled data to classify posts. However based on the unstructured nature considering only labelled data and NLP would not be sufficient to identify issue posts. The result of this research has led to the formulation of a gap. As a solution to this problem, the research proposes an enhanced combined system that aims to classify and predict issue posts.

The proposed system combines the use of textual context for feature extraction along with the combined machine learning model and validity algorithm to accurately identify issues posts. The system will take Stack Overflow data to identify issues. Necessary features related to the question will be generated using sentiment and word similarity score. Then the extracted features will be used on the combined model. Once the classification process is completed a post validity calculation is conducted by considering features such as user reputation, upvotes, downvotes and post score to measure the validity of the post. These extracted issue posts are then predicted and presented to the end user.

Keywords : Natural Language Processing, Sentimental Analysis, Similarity Measures, Neural Networks, Machine Learning, Clustering and Classification