INFORMATICS INSTITUTE OF TECHNOLOGY
In Collaboration with
UNIVERSITY OF WESTMINSTER

Crypto and Forex Trading Scam Tweets Detection System

Final Thesis Report by

Mr. Seyed Ruzaik

Supervised by

Ms. Sulochana Rupasinghe

Submitted in partial fulfilment of the requirements for the BSc (Hons) in Computer Science degree at the University of Westminster.

April 2023

Abstract

Spam detection on Twitter using text classification is a well-developed topic of NLP. A lot of research is already being undertaken in this field, and the outcomes of that research have made this field fully developed given the existing resource constraints. However, because to the complexity and constraints of numerous processes in the systematic NLP techniques, detecting crypto/forex trading scams is one of the specialist topics that hasn't been addressed previously.

This research provides an overview of how to detect potential scam tweets for crypto/forex trading using an ensemble technique for a binary classification model. Firstly, beginning with pre-processing the tweets to clean and format them for modeling. Then, the dataset was divided into training and testing sets using train-test-split technique, and a Random Forest Classifier model was fitted using the training set. For each tweet in the dataset, the aggregated predictions from each tree result in a final prediction.

A benchmark analysis is performed utilizing multiple algorithm approaches such as Support Vector Machine (SVM), Logistic Regression (LR), KNeighborsClassifier (KN), Random Forest Classifier (RF) etc. So that to get the most suitable and the classification method with the lowest possible error rate for the implementation of the prototype. According on the findings of the benchmark analysis, the Random Forest Classifier Method surpasses other approaches by giving an overall-accuracy-result of 98.7% on unseen data. Therefore, the Random Forest Classifier technique was taken into account considering how well it performed with the predictions.

Keywords - Binary Classification, Machine Learning, Natural Language Process, Twitter, Tweets, Crypto, Forex Trading, Bitcoin, Random Forest Classifier, Scam Detection, Ensemble Approach

Subject Descriptors

Computing Methodologies→Natural Language Processing→Text Classification→Ensemble approaches→Random Forest Classifier