Adedoyin Alaba
19221436

Detecting spam campaigns on Twitter using machine learning approach

Configuration Manual

Twitter Spam Detection with a focus on Election

Social networking sites have become very popular in recent years. Users use them to find new friends, updates their existing friends with their latest thoughts and activities. Among these sites, Twitter is the fastest growing site. Its popularity also attracts many spammers to infiltrate legitimate users accounts with a large amount of spam messages. In this paper, we discuss some user-based and content-based features that are different between spammers and legitimate users. Then, we use these features to facilitate spam detection. Using the API methods provided by Twitter, we crawled active Twitter users, their followers/following information and their most recent 100 tweets. Then, we evaluated
our detection scheme based on the suggested user and content-based features. Our results show that among the four classifiers we evaluated, the random Forest
classifier produces the best results. Our spam detector can achieve 95.7% precision and 95.7% F-measure using the Random Forest classifier.

File name definition

`src` : is a folder that contains all the utilities functions in python scripts that was used to preprocess the data as well as explore the data

`notebooks` : contains ipynb files that explores the data and clean up the data and build the model

`data` : is used to store some of the data that have been used

`images` : contains the images that was generated during the modelling & Evaluation stage of the peoject

`models` : contains models that was built for this project

`src` : contains the scripts files

- `account_checker.py` : for checking the autheticity of a twitter account
- `classifier_spam` : a python script to classify whether a tweet is spam or not
- `confusion_matrix` : to plot the confusion matrix of the model
- `data_extractor` : used to get data from twitter
- `modelling` : to build the models and visualize results
- `predict_account_type` : to predict different account type
- `screenNameCollector` : a script to collect screen Name from list of Twitter users

To Recreate this Project

User Needs a Python Environment installed on local Machine

1. User needs to clone this repo and change directory to the newly created folder
2. run `pip install -r requirements.txt`
3. run `python src/modelling.py`