Suspicious Content and Profile Identification Based on Quantifying Data on Twitter

by

Surendra Singh Gangwar

Roll. No.: 2016IPG-107

ABV–INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT GWALIOR (M.P.), INDIA

Motivation

Online Social networking is the medium of exchanging activities, entertainment and information. Although the social network provides huge benefits to the persons but at the same time also harms with different malicious social activities.This causes our society considerable economic loss and national security even threatens. Because of URL shorteners, common and informal languages and abbreviations used on social networking sites filtering out the malicious content becomes a challenging problem.

Key related research

For spam detection numerous methods utilized by researchers.

- User Based Techniques: A user's account includes important information like No. of followers,No. of Following, No. of mentions, Tweets creation time.
- Content Based Techniques: A tweet content has some crucial information like Number of hashtags in comparison to total word count, Users mentioned in a tweet, Number of URLs, count of numerals etc.
- Relation Based Features: It uses a connection degree whether a person mentioned a direct friend in a tweet or a mutual friend.

Review of key related research

This section discusses the recent works in the field of spam detection on twitter andtheir shortcomings.

- Dangkesee and Puntheeranurak [1] perform an adaptive classification for spam detection. They used spam world filter and url filter using blacklisted urls and found proposed model outperformed spam world filters by com-paring Accuracy, precision, recall, f1-score.
- Raj et al. [2] proposed multiple machine learning algorithms to classify tweet content. In KNN(92%), Decision Tree classifier(90%), Random Forest Classifier(93%),Naive Bayes classifier(69%), Random forest outperformed.
- Song et al. [3] presented Bagging, SVM, J48, BayesNet with relation based features by creating graph b/w users they used distance and connectivity b/w users.Finally Bagging outperformed with 94.6% true positive and 6.5% false positive.

Review of key related research

- Alom et al. [4] conducted CNN with tweet text and with both tweet text and meta-data features. This method out performed other deep learning methods so they suggest to use this in other social networks like linkedin and facebook.
- Mateen et al. [5] proposed a hybrid approach for spam detection in which they used different combinations of approaches like content based and graph based feature and user based with graph based features. Applies J48, decorate and naive bayes classifier with these features. Accuracy for content and graph based features achieved 90% and for user and graph based features achieved 92%. They also find correlated features and remove features with higher correlation.

Objectives

- Try to develop a system where we can pass real time tweets to classify them as spam or non-spam and user as legit or spam.
- Compare different methods and their performances on the model using different preprocessing techniques.
- Using different combinations of content, user and relation based features with different ensemble based learning.
- Extend this work to generate the more efficient model by bringing parameter tuning into consideration.

Novelty of the proposal

- Our proposed model uses the combination of different features like User-Relation, Content-Relation, User, Combination of all three User, Content and Relation features. Further we will also implement URL filter with safe browsing.
- We will aim to apply our proposed model with some specific constraints like language and trending topic like covid-19.
- Extend this work to generate the more efficient model by bringing parameter tuning into consideration.

Dataset Description

- The proposed work utilizes the tweets fetched using twitter developer API. Twitter allows its users to fetch twitter data using tweepy library. We fetched 2000 latest tweets which consists of many features like timestamp, tweet text, username, hashtags, followers count, following count, number of mentions, word count, is retweet etc.
- All of these features are categorized in content based features and user based features.Later three different datasets will be created by combining user-content features, user-relation features, user-content-relation features. These dataset will be utilized for our proposed model.

Methodology

Below architecture shows the various steps which would be involved while developing our Spam detection Model.

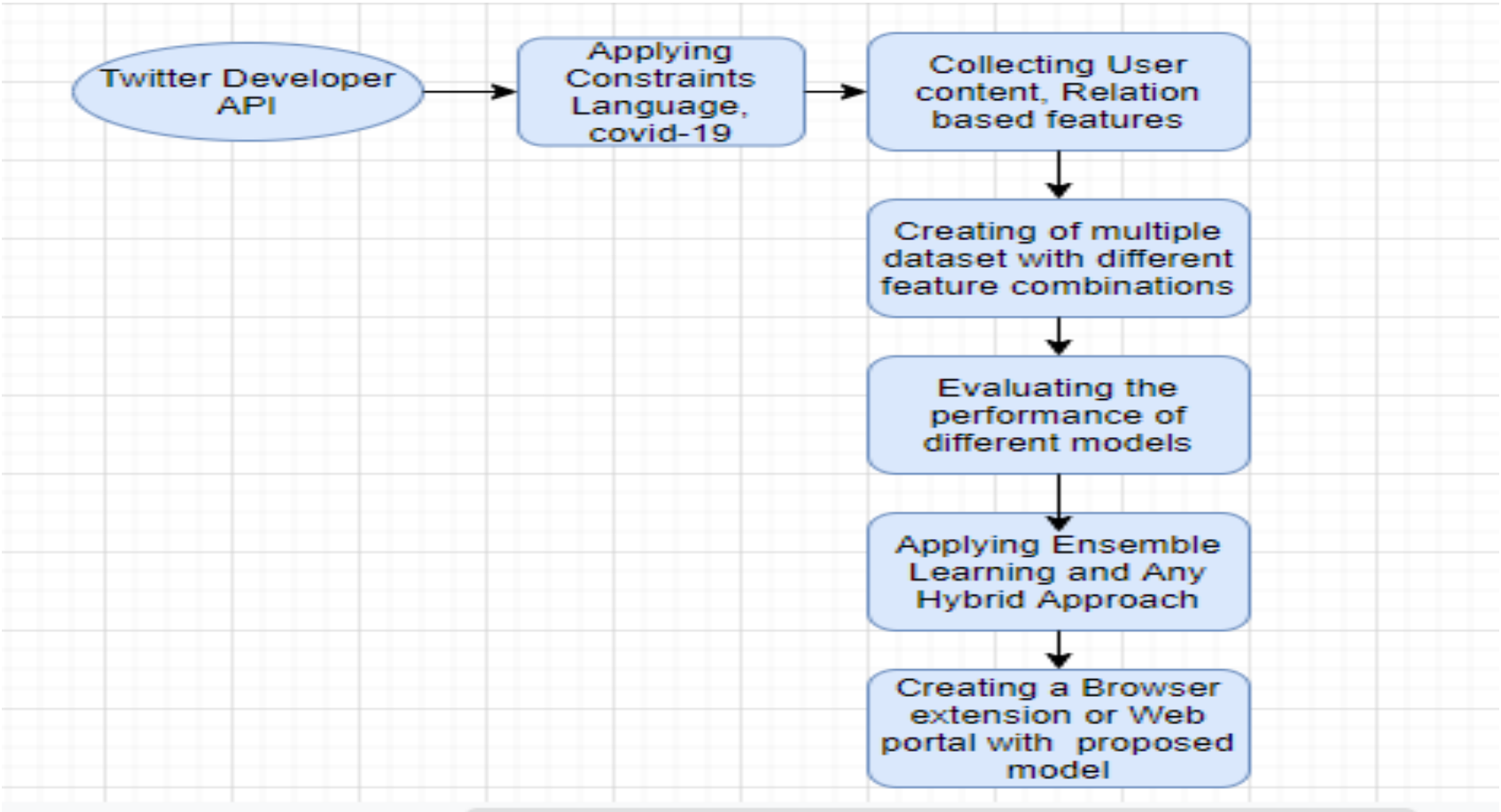

Figure: Methodology

Methodology

Methodology represents different steps involved in the development of the model.

- Tweet Collection : First the framework fetches tweets utilizing twitter dev API.

In [15]:
```
"""
INPUTS:
    consumer_key, consumer_secret, access_token, access_token_secret: codes
    telling twitter that we are authorized to access this data
    hashtag_phrase: the combination of hashtags to search for
OUTPUTS:
    none, simply save the tweet info to a spreadsheet
"""
def search_for_hashtags(consumer_key, consumer_secret, access_token, access_token_secret, hashtag_phrase):

    #create authentication for accessing Twitter
    auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)

    #initialize Tweepy API
    api = tweepy.API(auth)

    #get the name of the spreadsheet we will write to
    fname = '_'.join(re.findall(r"#(\w+)", hashtag_phrase))

    #open the spreadsheet we will write to
    with open('%s.csv' % (fname), 'w', encoding='utf8') as file:
    #with open('%s.csv' % (fname), 'wb') as file:

        w = csv.writer(file)

        #write header row to spreadsheet
        w.writerow(['timestamp', 'tweet_text', 'username', 'all_hashtags', 'followers_count'])

        #for each tweet matching our hashtags, write relevant info to the spreadsheet
        for tweet in tweepy.Cursor(api.search, q=hashtag_phrase+' -filter:retweets', \
                                   lang="en", tweet_mode='extended').items(1500):
            w.writerow([tweet.created_at, tweet.full_text.replace('\n',' ').encode('utf-8'), tweet.user.screen_name.encode('utf-
```

In [16]:
```
consumer_key = input('Consumer Key')
consumer_secret = input('Consumer Secret')
access_token = input('Access Token')
access_token_secret = input('Access Token Secret')

hashtag_phrase = input('Hashtag Phrase')

if __name__ == '__main__':
    search_for_hashtags(consumer_key, consumer_secret, access_token, access_token_secret, hashtag_phrase)
```

Figure: Twitter API

Methodology

- Spam labelling : Initially, All of the tweets are unlabeled, They need to be labeled as spam or non-spam for training purposes.
- Evaluating the performance of different Machine learning Models: In this step we will train KNN, Naive Bayes classifier, Decision Tree Classifier and various Ensemble methods like Bagging and Boosting with the prepared dataset and will evaluate the performance of these models.

Methodology

- Feature extraction : There may be some crucial features which play an important role for classifying the tweets. So in these steps these features will be extracted from the dataset.
- Hybrid Model: Finally we will apply hybrid model with different datasets created with combination of different feature types. We will compare the performance of our Hybrid model with different machine learning models and ensemble methods.

Plan of action

Below chart represents time ranges of for different steps involved in thesis.

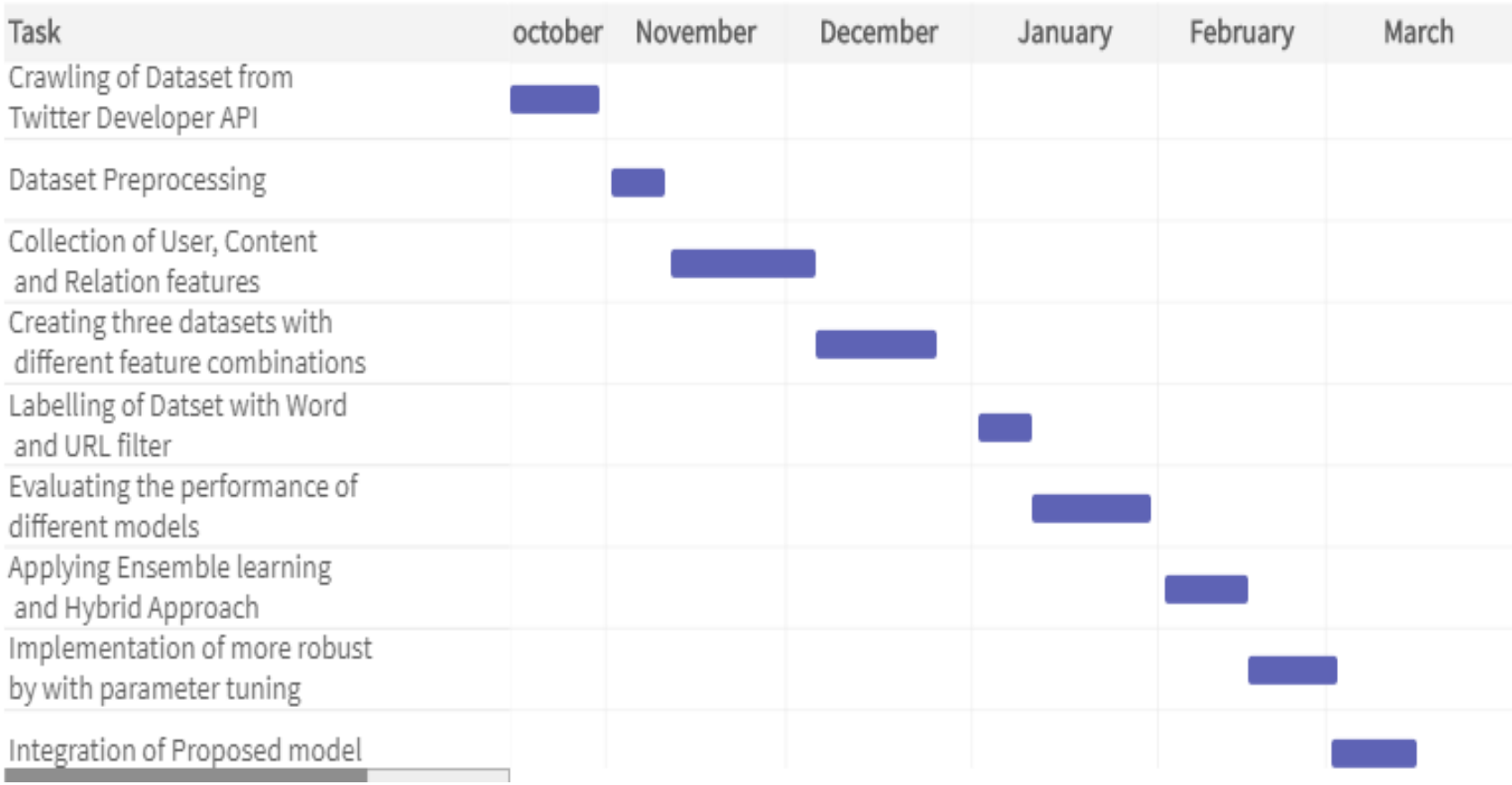

Figure: Gantt Chart

Expected research outcome

The expected outcomes during the course of research have been discussed below:

- Evaluating the performance of various techniques on datasets with different feature combinations.
- To deliver a robust algorithm for detecting spam tweets with a great accuracy, precision, recall, f1-score.
- Use parameter tuning to come up with even more advanced Architecture for developing robust model.

References I

Thayakorn Dangkesee and Sutheera Puntheeranurak.
Adaptive classification for spam detection on twitter with specific data.
In *2017 21st International Computer Science and Engineering Conference (ICSEC)*, pages 1–4. IEEE, 2017.

R Jeberson Retna Raj, Senduru Srinivasulu, and Aldrin Ashutosh.
A multi-classifier framework for detecting spam and fake spam messages in twitter.
In *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pages 266–270. IEEE, 2020.

References II


Jonghyuk Song, Sangho Lee, and Jong Kim.
Spam filtering in twitter using sender-receiver relationship.
In *International workshop on recent advances in intrusion detection*, pages 301–317. Springer, 2011.


Zulfikar Alom, Barbara Carminati, and Elena Ferrari.
A deep learning model for twitter spam detection.
Online Social Networks and Media, 18:100079, 2020.


Malik Mateen, Muhammad Azhar Iqbal, Muhammad Aleem, and Muhammad Arshad Islam.
A hybrid approach for spam detection for twitter.
In *2017 14th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pages 466–471. IEEE, 2017.