جامعة بيرزيت

BIRZEIT UNIVERSITY

Electrical and Computer Engineering Department ENCS434

Artificial Intelligence

ENEE4103

Programming Project 2

"Automatic Tweet Spam Detection"

Partner's Name :

Lama Hasan 1182525

Baraa Amer 1182595

Yahya Alqarout 1171257

Instructor:

Dr. Aziz Qaroush

Date :20/1/2022

ABSTRACT

The aim of the project is to become familiar with machine learning techniques. The idea of This project is how to build practical applications like Automatic Tweet Spam Detection using machine learning techniques , we have read and understood how spammers try to post tweets and based on that we extracted the features which will help to detect which the tweet is spam or not . We built four different models, including: Decision tree, Naïve Byes, Neural Network and Random Forest classifier and we used the Python language and its libraries like Scikit learn related to machine learning algorithms.

Contents

Table of figure

INTRODUCTION

Machine learning is a specific branch of AI. Its concept is founded on the ability of machines to learn by themselves, instead of humans teaching computers everything they need to know. This allows machines to imitate and adapt human-like behavior, and There are three types of machine learning:

1. **supervised learning**
2. **unsupervised learning**
3. **reinforcement learning**

Supervised learning

Supervised machine learning uses an algorithm with the correct label for a dataset of examples. Later, it uses its model to label new input by finding similarities in characteristics (indicators, predictors) between the examples and new input. Supervised learning systems need to know the labels that you want upfront. In general, companies prefer to phrase the solution to a problem by means of supervised learning.

Unsupervised learning

Unsupervised learning does not include pre-specified label input. It simply means that new input is grouped in terms of similarities at random. A computer analyses certain input and returns it as clustered groups. There are many different characteristics it can use to define similarities and thus many different ways to cluster the same input. This method is relevant when you want to let a machine specify patterns and groups, but do not know upfront what you are looking for (labels).

Reinforcement learning

Reinforcement learning refers to an algorithm that learns to react to its environment. This is the most complex form of machine learning. In this concept, a machine (such as a robot) performs an action in its environment. This is interpreted into a reward and representation of the state, which are fed back to the machine. This form of learning is closest to how a human brain works and to how a human develops.

RELATED WORK

Techniques to Detect Spammers in Twitter- A Survey.

(Verma, Divya, & Sofat, 2014)

With the great popularity of Online Social Networking sites (OSNs) cybercrimes become more common and easy for spammers. According to the paper, spammers are users who misuse the OSNs for profit purpose by stealing other users sensitive data using the attention of advertising products or harmful URLs, the paper explains a review for previous studies about methods to detect spammers.

The methodology used in this paper chose 60 related papers after reading its tittle and abstract, Irrelevant topics are excluded for this papers, so finally a total of 21 papers have been selected, these papers have been categorized based on the features used to detect spammers .the survey described the methodology, data set, features to detect spammers and accuracy for the techniques used, so any researcher can easily benefits from the survey.

There are many security and privacy issues happened for user's sensitive data, these attacks are briefed in Viruses, Phishing attacks, Phishing attacks, Sybil (fake) attack , Social bots ,and Clone and identity theft attacks. spammers categorized as phishers, fake Users and promoters. Also, There are many purposes for spammers Disseminate pornography, Spread viruses, Phishing attacks and Compromise system reputation.

In the next section, the paper shows detailed information about tweeter as a social media platform, it also defined Threats on Twitter as Threats on Twitter, Malware downloads and Twitter bots.

The paper mentioned Features could be used to easily detect spammers tweet, these features can be user based, content based or both, an example of feature to detect spam numbers of followers (spammers have less number of followers), age of account (new created accounts), idle hours (spammers keep sending messages so they have less idle hours), number of hashtages (#) and mentions (spammers use maximum @usernames), retweets (spammers use maximum @RT) and HTTP links (number of www or http://), and many others mentioned in the paper.

The paper showed the summary of the papers reviewed regarding the detection of spammers in Twitter, it compared between these papers based on data set and methodologies such that SVM, Decision Tree, Naive Bayesian, and Random Forest by calculating the accuracy for these methodologies.

Twitter spam detection: Survey of new approaches and comparative study.

(Wu, Wen, & Zhou, 2017)

The Twitter application is one of the most important social media in this modern , and it is similar to the Facebook application, which enables people to communicate on the Internet and share the opinions between them. As there are more than a billion users on Twitter, as a result of this large number of tweets, some users have appeared who publish spam (unwanted) tweets in various fields. Based on previous statistics, every person who used social media encountered spam of some kind.

There are many forms that social media users are exposed as in the following:

Viruses :spammers use the twitter as a platform to spread malicious data in the system of users.

Phishing attacks : Sensitive user information is obtained by impersonating a genuine user.

Spammers - send spam messages or post spam tweets to the users of social networks.

fake attack - attacker obtains multiple fake identities and pretends to be genuine in the system in order to harm the reputation of honest users in the network.

Clone and identity theft attacks: Hackers create a social media account with the same personal data as a genuine user to defraud their friends.

TYPES OF SPAMMERS:

Spammers are defined as malicious users who create fake accounts and information to pose a threat to the security and privacy of users and social media, so, They are categorized as:

Phishers: They pretend to be genuine users to get personal data of other genuine users

Fake Users: They are users who impersonate genuine people to deceive their friends

Promoters : They are the ones who send malicious links to advertisements or promotional links to sites that steal user data

Summers goals in general:

a) Disseminate pornography b) Spread viruses c) Phishing attacks d) Compromise system reputation

As we know Twitter uses a chirping bird as his emblem thereforeTwitter name. Users can access it for frequent exchange Information called "tweets" which are messages of up to 140Long characters that anyone can send or read.

Users share these tweets which may contain news,Reviews, photos, videos, links and messages. Below Standard terms used on Twitter and relevant to our work:

Tweets : A message on Twitter containing maximum length of 140 characters.

Followers & Followings [3]: Followers are the users who are following a particular user and followings are the users whom user follows.

Retweet: A tweet that has been reshared with all followers of a user.

Hashtages: twitter uses this # symbol so that users can target specific words or topics that are known and easily accessible to other users.

Mentions: Twitter will use the @ symbol in order to allow users to tag their friends.

List : Twitter provides a mechanism to list users you follow into groups

Direct Message [3]: private masseges between users.

URL : Spammers post a lot of these bad url in their tweets

Spam Words- Spammer's tweets mainly consist of spam words.

HTTP links- if tweets contain maximum number of www or http://, then they are posted by spammers.

Duplicate tweets- spammers tend to post duplicate tweets with different @usernames in tweets.

There are some Existing Methods for detections of spam profiles in twitter like :

1-Benevenuto et. al. [7] detected spammers on the basis of tweet content and user based features. Tweet content attributes usedare - number of hashtags per number of words in each tweet, number of URLs per word, number of words of each tweet, number of characters of each tweet, number of URLs in each tweet, number of hashtags in each tweet, number of numeric characters that appear in the text, number of users mentioned in each tweet, number of times the tweet has been retweeted. Fraction of tweets containing URLs, fraction of tweets that contains spam words, and average number of words that are hashtags on the tweets are the characteristics that differentiate spammers from non spammers. Dataset of 54 million users on Twitter has been crawled with 1065 users manually labelled as spammers and non-spammers. A supervised machine learning scheme i.e. SVM classifier has been used to distinguish between spammers and non

spammers. Detection accuracy of the system is 87.6% with only 3.6% non-spammers misclassified.

2-Twitter facilitates its users to report spam users to them by sending a message to "@spam". So Gee et. al. [12] utilized this feature and detected spam profiles using classification technique. Normal user profiles have been collected using Twitter API and spam profiles have been collected from "@spam" in Twitter. Collected data was represented in JSON then it was presented in matrix form using CSV format. Matrix has users as rows and features as columns. Then CSV files were trained using Naive Bayes algorithm with 27% error rate then SVM algorithm has been used with error rate of 10%. Spam profiles detection accuracy is 89.3%. Limitation of this approach is that not very technical features have been used for detection and precision is also less i.e. 89.3% so it has been suggested that aggressive deployment of any system should be done only if precision is more than 99%

During this research, it was found that different ways of detecting spam systems and each time some improvements are made to get a higher detection rate.

1. Since Twitter has millions of active users and this number is constantly increasing. And almost all the authors have used very small testing dataset to see the performance of their approach. So there is a need to increase the testing dataset to see the performance of any

approach.

2. Need to develop a multivariate model.

3. Need to develop a method that can detect all kinds of spammers.

4. Need to test the approaches on different combinations of spammers and non-spammers.

Through, the most widely used classifier like : SVM, Naïve Byes And we extracted auxiliary features to detect based on features for the user or for the content These systems were validated on a small data set have not been even tested with different combinations of spammers and non-spammers. Combination of features for detection of spammers has shown better performance in terms of accuracy, precision, recall etc. as compared to using only user based or content based features.

A Survey of Spam Detection Methods on Twitter

(Kara, 2017)

Twitter is one of the most popular social media platforms,the popularity of twitter attracts the attention of spammers for their malicious aims. structure of Twitter which is using “Trending Topics (TT)” and “Hashtag” let users track the topics they are interested in and lets real-time search systems and meme-tracking services mine real-time tweets to find.

The success of those services completely relies on filtering spammers from legitimate users, Most common definition of spam is unsolicited one those spammers share links within their tweets in order to spread advertises for their aims.

According to the report by Statista, most of Twitter users access Twitter via their mobile devices ,those users should more care cause spammers can access thiere sensitive data.

spammers tend to share shorten URLs in order to (1) overcome the character limitation defined by Twitter, and (2) manipulate spam filtering methods based on URL.

The Features of Twitter from article is Twitter lets accounts to “follow” other accounts , the relationship between users is bi-directional, Each user has a unique Twitter username, and users can post tweets that refer others and letting users create public or private lists and added users to these lists.

How Twitter Deals with Spam:
both manual and automated services
The manual is outdated is lets users report or mentioning spammers to the official “@spam” account.

The automated is using factors posting duplicate messages, following/unfollowing large number of accounts in a short time period, having large number of spam complaints filed against the account, aggressively liking, following, and retweeting, posting malicious links, posting tweets which mainly consist of links instead of also posting personal updates and posting unrelated tweets to a trending topic to determine what conduct is considered to be spamming.

From the article the Features of twitter spam detection is Account-based features, since some of these features are user-controlled, they are useless in term of spam detection, tweet-based features, by analyzing their tweets , Spammers tend to use links , lots of mentions and lots of hashtags,the number of likes and retweets their tweets and relationship between the tweet's sender and receiver.

Finally twitter spam detection methods is Account-based Spam Detectıon Methods, Tweet-based Spam Detection Methods, Graph-based Spam Detection Methods and Hybrid Spam Detection Methods.

PROPOSED WORK :

We build the project and program it based on the following steps :

1- pre-processing :

This step the first thing we must to do ,it is very important for this project ,it is essential for make everything is ready for extract features .for example , we fill the misleading values in dataset in our project we fill the misleading values with the average value in the column because the misleading values causes the problems in the training model. Also we extract some characters in the tweet and the location in order to build the high performance classifier.

2- extract features :

As we know, The dataset contains useful and non-useful information.so we extract the features the which affect to the to the classifier strongly. And ignore the other non useful information we can extract several features from one column in raw data, in our project we found the extract features as the following:

2.1 Number of hashtags: According to our searches, we found that the spammers use a lot of hashtags to attract attention through followers.

2.2 Number of mention: According to our searches, we found that the spammers also use a lot of mention to attract attention through followers.

2.3 Special content in Tweet: we found that the spammers use special URL in their tweets as this (https://t.co/ ….) this used a lot in tweets of spammers.

2.4 Following and Followers: we found the relationship between following and follower and Type of tweet. We found f the following of spammers greater than its followers then the type of sweet is spam s, we extract feature compare between following and followers.

2.5 Content URL in tweet: we found that the spammers use a url a lot in order to spread their malicious links and harmful content to people.

2.6 Action: we found that if the tweet have a large number of action this often quality**.**

2.7 Is retweet: we found that if the tweet is retweet then in most cases is spam .

3- Models Training:

In our project we implement three different classifiers as the following:

3.1 Decision Tree:

This used as supervised machine learning and with application which have the output labels, It aims to create a model capable of predicting the targeted variable. It uses simple decision rules inferred from the data features

3.2 Naïve Byes classifier:

It is constructing classifiers: models that assign class labels to problem instances , a probabilistic machine learning model, This algorithm based on the byes theorem with strong independence assumptions between the features.

3.3 Neural Network:

This used as supervised machine learning and with application which have the output labels, are comprised of a node layers, containing an input layer, one or more hidden layers each node, or artificial neuron, connects to another and has an associated weight and threshold.

3.4 Random Forest

Is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. ... For regression tasks, the mean or average prediction of the individual trees is returned.

*In all above algorithms, we can split the dataset to 80% of data to training the model to be able to predict the tweets in new data and 20% of data to test the model and check the performance of the model like accuracy, precision and recall.

EXPERIMENT AND RESULT

Notice that for all classifier we have split the data set for 80% training and 20 % testing.

In Following figure is the accuracy and classification report for the decision tree, We noticed that the accuracy is 99% which is very high percent and almost correctly to detect the tweet is spam or not.

```
the Accuracy for my model use decision tree is :
% 99.20634920634922
DecisionTree report
              precision    recall  f1-score   support

     Quality       0.99      0.99      0.99      1272
        Spam       0.99      0.99      0.99      1122

    accuracy                           0.99      2394
   macro avg       0.99      0.99      0.99      2394
weighted avg       0.99      0.99      0.99      2394
```

Figure 1:Decision Tree model

In Following figure is the accuracy and classification report for the Naive Bayes model, we noticed that the accuracy is 85% which is less one compared with other models.

```
the Accuracy for my model use Naive Bayes is :
% 85.04594820384294
 naiveBayes report
              precision    recall  f1-score   support

     Quality       0.78      0.99      0.87      1235
        Spam       0.99      0.70      0.82      1159

    accuracy                           0.85      2394
   macro avg       0.88      0.85      0.85      2394
weighted avg       0.88      0.85      0.85      2394
```

Figure 2:Naive Bayes model

In Following figure is the accuracy and classification report for the Neural Network model, we noticed that the accuracy is 93% which is good accuracy.

```
the Accuracy for my model use neuralnetwork is :
% 93.35839598997494
neuralnetwork report
              precision    recall  f1-score   support

     Quality       0.90      0.98      0.94      1251
        Spam       0.97      0.89      0.93      1143

    accuracy                           0.93      2394
   macro avg       0.94      0.93      0.93      2394
weighted avg       0.94      0.93      0.93      2394
```

Figure 3:Neural Network model

In Following figure is the accuracy and classification report for the Random forest model, we noticed that the accuracy is 99% which is very good accuracy to detect the tweet is spam or quality.

```
the Accuracy for my model use Random forest is :
% 99.5405179615706
Random forest report
              precision    recall  f1-score   support

     Quality       0.99      1.00      1.00      1226
        Spam       1.00      0.99      1.00      1168

    accuracy                           1.00      2394
   macro avg       1.00      1.00      1.00      2394
weighted avg       1.00      1.00      1.00      2394
```

Figure 4:Random Forest model

In the following chart show the difference between the accuracy for each model .

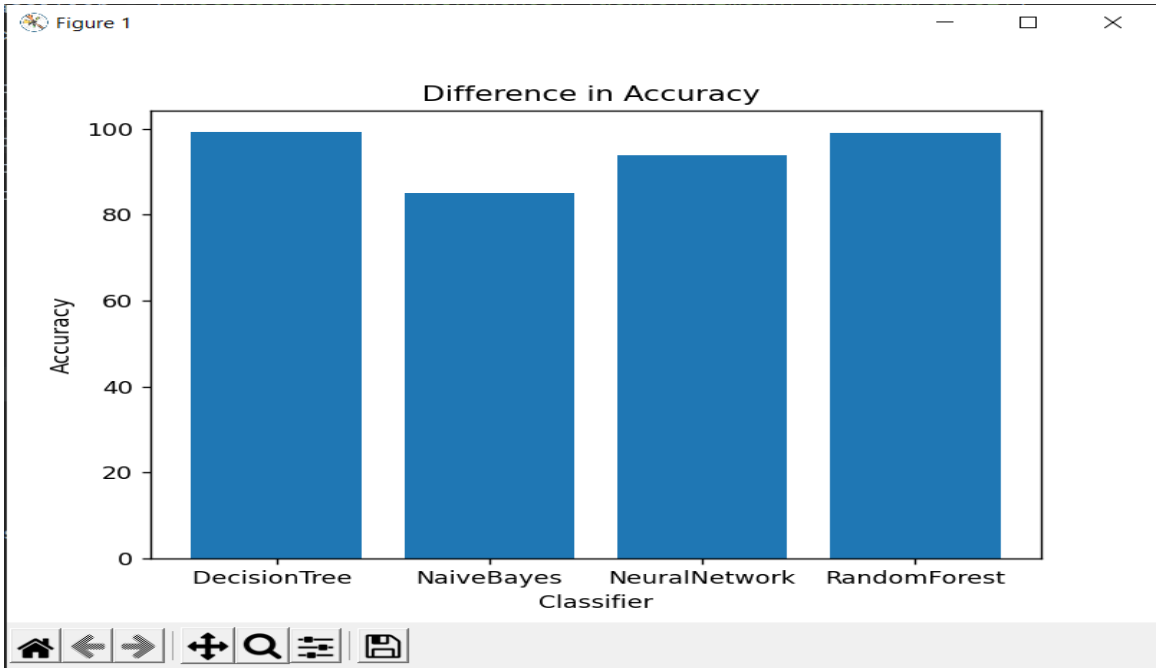

Figure 5: Difference between model

CONCLUSION

In this project we have learned a lot of new and useful topics in the machine learning, so that we were able to build a model in different ways in order to detect the tweet whether it is a spam or not. We conclude that the python language and its libraries form a strong and important part in building the Ai Applications We built three Classifiers and we saw the difference between them through the results like Accuracy, Precision and Recall and the results were excellent.

REFERENCES

1. https://www.cbi.eu/market-information/outsourcing-itobpo/machine-learning-artificial-intelligence?gclid=EAIaIQobChMIqsLH6qbG9QIVELh3Ch3HLAP3EAAYASAAEgKm9PD_BwE on 18/1/2022 at 10:00 pm
2. https://www.google.com/search?q=random+forest&oq=Random+Forest&aqs=chrome.0.0i512l4j69i59j69i60l2j69i61.352j0j7&sourceid=chrome&ie=UTF-8 on 18/1/2022 at 11:00 pm

Bibliography

Kara, A. T. (2017, March). A Survey of Spam Detection Methods on Twitter.

Verma, M., Divya, & Sofat, S. (2014). Techniques to Detect Spammers in Twitter- A Survey.

Wu, T., Wen, S., & Zhou, Y. X. (2017). Twitter spam detection: Survey of new approaches and comparative study.