**TITLE OF THE PROJECT**

# FAKE ACCOUNT PREDICTION

A Project Report

Guided by

Dr.Syed Khasim

T DEVI SAI JNANESWAR VUNDAVILLI - 20BCD7274

R ESWARA VENKATA SAI - 20BCI7230

UJJWAL REDDY KS - 20BCI7203

YASWANTH - 20BCI7331



VIT-AP UNIVERSITY

AMARAVATHI NOVEMBER-2022

# DATA ANALYTICS PROJECT

**FAKE ACCOUNT PREDICTION**

**Abstract**:

Information is created and spread by millions of people in online social networks. People can follow trending topics and events online with the use of search engines and data mining tools. However, the widespread usage of social media also facilitates the exploitation of social media by bad users, also known as spammers, who fill social media with useless content. A classifier is required to find social spammers in order to resolve this issue. Based on user data and content, one reliable method for spammer detection. However, social spammers are cunning and adept at tricking the system by constantly updating their information and contents. First off, social spammers constantly alter their behaviours in an effort to trick detection software. Second, spammers will make every effort to blend in and influence others. Existing solutions struggle to respond to social spammers effectively and efficiently due to their dynamic behaviour. In this study, we provide a logistic regression model that takes into account user behaviour and content qualities in social networks. Analysis of user attributes is done to automatically distinguish between spammers and non-spammers. Results from experiments using data from Twitter demonstrate the usefulness and efficiency of the suggested approach.

**Key** **words —** Social network · Social spammer · Classifier · Logistic regression

**Packages:**

1.sklearn.datasets 2. matplotlib

1. sklearn.metrics
2. seaborn
3. pandas

6.DT

7.scales

**Problem Statement:**

. There are differences between social spammers and normal users in attributes and behaviors. Normal users will be more active than social spammers, which is reflected in their content attributes and user information. Social spammers disguise themselves as normal users generally, but when they launch attacks, their behavior patterns will be abnormal. So it is possible to take advantage of those factors which differentiate social spammers from normal users to build a classifier for spammer detection. The rest of the paper is organized as follows. The next Section introduce the background and related work of spammer detection in social network, several existing methods are presented. In Sect. 3, crawling strategy and labeled collection build method are depicted in detail.Then a classifier based on logistic regression model considering attributes and behaviors is introduced. Section 4 selects several evaluation metrics and makes comparison experiments using crawled data with existing methods, the ability to distinguish spammers and non-spammers is investigated. Section 5 makes conclusions and orientation of the future work.

**INTRODUCTION:**

Online social networks, like Facebook and Twitter, are increasingly used in many fields such as education, marketing and politics. While social network has become important platform for information diffusion and communication, it has also become disgraceful for social spammers who publish and propagate junk information. The flood of social spammers seriously influences the experience of normal accounts, detecting social spammers can make a great contribution to the development of online social network. It will promote the user experience and healthy use of the whole online social network.

Traditional spammer detecting methods become less effective due to the fast evaluation of social spammers. Social spammers try to disguise themselves as normal users by changing information and dynamic patterns to cheat the system. In order to detect social spammers effectively, it is necessary to build a new model to take new characters into account. There are differences between social spammers and normal users in attributes and behaviours. Normal users will be more active than social spammers, which is reflected in their content attributes and user information. Social spammers disguise themselves as normal users generally, but when they launch attacks, their behaviour patterns will be abnormal. So, it is possible to take advantage of those factors which differentiate social spammers from normal users to build a classifier for spammer detection.

**Literature review:**

Online social networks have become the preferred form of communication not only between friends and family, but also in business affairs. Friends and family members will share the life information and communicate with each other in online social networks. In order to increase exposure of a brand, lots of traditional shops build their accounts online to introduce their products and hold business campaign. It will attract people to participate in the activity and discuss about it, all the behaviours will make contributions to information diffusion. Social networks are important platforms for information diffusion and communication.

**Our main objectives of this project are:**

* Visualize the growth of Fake Accounts • Analyse the Ratio According to the time series;
* and calculate the Accuracy.
* Additional information on how the Fake Account is detected; **Description of Dataset:**

Sina Weibo is a Chinese microblogging (weibo) website. Akin to a hybrid of Facebook and Twitter, it is one of the most popular sites in China. A great deal of active users update their messages, repost interesting messages and establish friendships in Sina Weibo. We can also analyze users in Sina Weibo, the attributes of the users in Sina Weibo are similar to the attributes in Twitter.

**Methodology:**

**Initial Analysis**: Understand the dataset given. Look through its structure, identify the datatypes of various columns and get a basic idea of the dataset to proceed further.

**Data Cleaning**: Look out for visible data quality issues and rectify them. Check for blanks, duplicate data and convert certain columns to required datatypes.

**Exploratory Data Analysis**: Making use of Python, carry out various EDA operations like Univariate and Segmented Univariate analysis and come up with intuitive insights of the Supply-Demand problem.

**Visualization:** Make use of R libraries & package to plot various graphs with proper aesthetics and geometry, clearly displaying important insights.

**Proposed system:**

We proposed that we will build a data visualization project with ggplot2 using R and its libraries. Analyse various parameters like

1. Trips by the hours in a day’
2. Trips during months in a year.

In order to evaluate performance of our approach in spammer detection on online social networks, we need a labeled collection, in which users are pre-classified into spammers and non-spammers. To the best of our knowledge, no appropriate collection is publicly available, so we need to build one. we represent the method to build a labeled collection. Then we identify user’s attributes and discuss the process to manually classify a set of spammers and non-spammers in a method Attribute and Behavior Logistic Regression (ABLR), based on Logistic Regression Model to detect spammers.

# 1.Importing the Essential Packages

In the first step of our Data Analytics project, we will import the essential packages that we will use in this uber data analysis project. Some of the ***important libraries of Python*** that we will use are –

* **sklearn.datasets**

This is the backbone of this project. ggplot2 is the most popular data visualization library that is most widely used for creating aesthetic visualization plots.

* **matplotlib**

This is more of an add-on to our main ggplot2 library. With this, we can create better create extra themes and scales with the mainstream ggplot2 package.

* **sklearn.metrics**

This package is the lingua franca of data manipulation in python.

* **pandas**

This package will help you to tidy your data. The basic principle of tidyr is to tidy the columns where each variable is present in a column, each observation is represented by a row and each value depicts a cell.

* **Numpy**

With the help of this package, we will be able to interface with the ***javascript*** Library called – Data tables.

* **Seaborn**

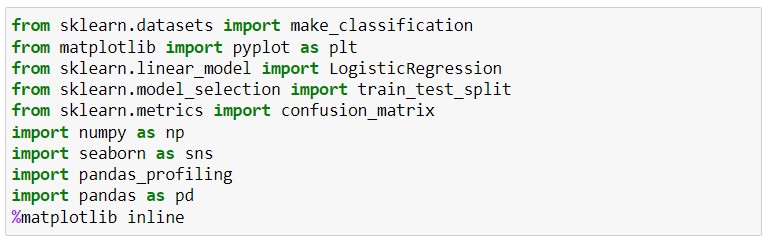
With the help of graphical scales, we can automatically map the data to the correct scales with well-placed axes and legends. (They take your data and turn it into something that you can see, like size, colour, position or shape.)

* **Classifier Based on Logistic Regression Model**

Logistic regression model is a classifier that focus on the binary classification problem in which result can take on only two values, 0 and

1 . So we can build a spammer classifier for online social network, then inputs are some features of a user, and output maybe 1 if it is a spammer user, or 0 otherwise. 0 is also called negative class, and 1 is called

positive class. The ABLR method is based on this model. Logistic regression is based on a hypothesis function hθ(x)



## 2. Reading the Data into their designated variables we will

create a vector of our colours that are in our plotting functions. You can also select your own colours that you want.

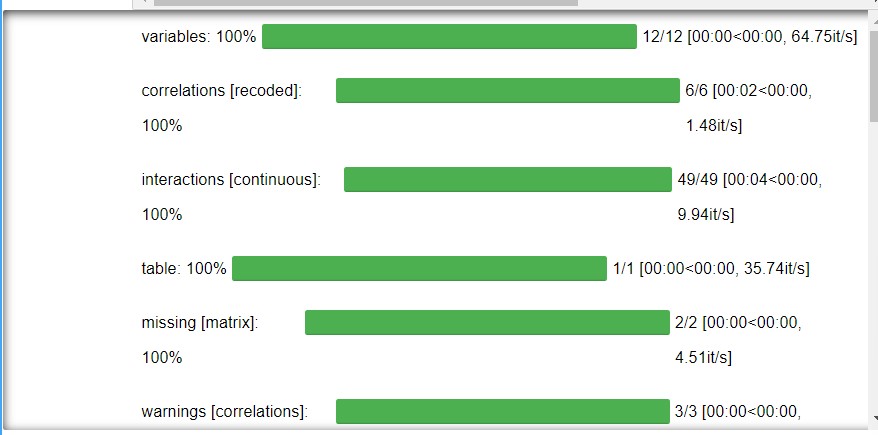


## 3. Viewing dataset

we will read the several csv files that contain the data from April 2014 to

September 2014. We will store these in corresponding data frames like apr\_data,

may\_data, etc. After we have read the files, we will combine all of this data into a single data frame called ‘data\_2014’.



**Plotting the Fake Accounts Using External URL’S**

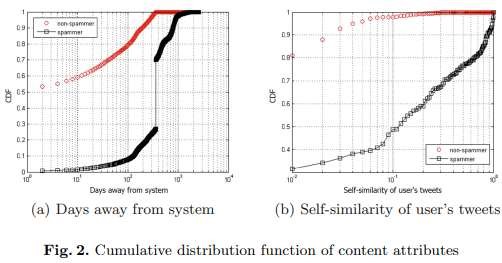
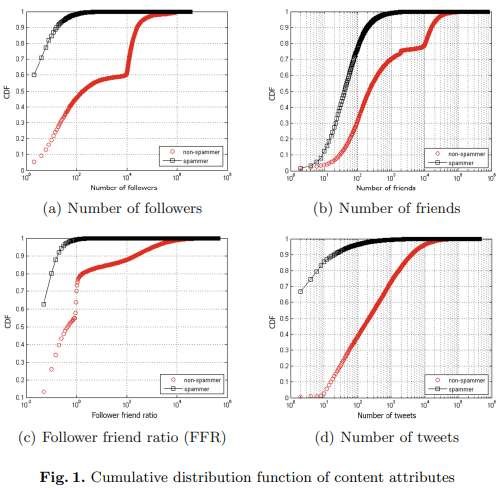
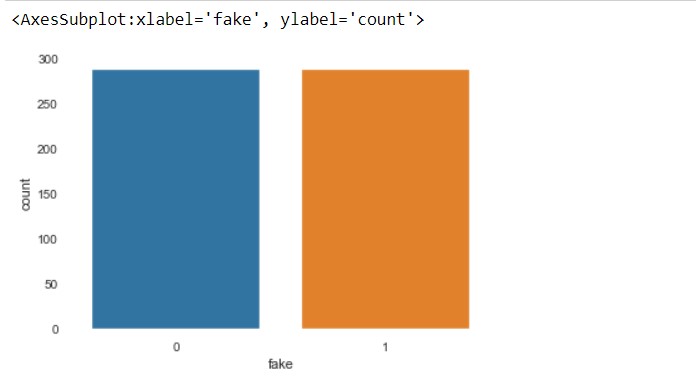
Here we will be using the ggplot function to plot the number of followers that the users had made in a day. We will also use dplyr to aggregate our data. In the

resulting visualizations, we can understand how the number of users create fake

account throughout the day. We observe that the number of fake accounts are

created in external URL’s

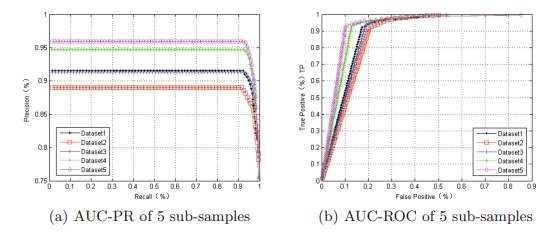




**Summary:**

Since users and tweets have a 1-to-N relationship, we need to figure a way for machine learning to map the tweet documents into an account classification scheme. One approach used in this research is to concatenate user tweets into a single document. This research has done this in two ways described in Section III-C. Meanwhile, the machine learning process of the whole consists of the following steps: data acquisition, data cleaning, feature extraction, modeling, and evaluation.

**Conclusion:**



In this paper we have presented a machine learning pipeline for detecting fake accounts in online social networks. Rather than making a prediction for each individual account, our system classifies clusters of fake accounts to determine whether they have been created by the same actor. Our evaluation on both insample and out-of-sample data showed strong performance, and we have used the system in production to find and restrict more than 250,000 accounts.In this paper we have presented a machine learning pipeline for detecting fake accounts in online social networks. Rather than making a prediction for each individual account, our system classifies clusters of fake accounts to determine whether they have been created by the same actor. Our evaluation on both in-sample and outof-sample data showed strong performance, and we have used the system in production to find and restrict more than 250,000 accounts.

**Future Scope:**

Social spammers are tricky and change their strategies continually to deceive social network system. To solve this problem, a method considering both content attributes and behavior attributes based on logistic regression. A training dataset is built up through crawling from twitter API. Several attributes are taken into account by analyzing cumulative distribution function of labeled datasets. Those attributes that can separate spammers from non-spammers are used to establish our classifier. Experimental results show that our proposed method is effective and efficient comparing with existing social spammer detecting method. This work suggests some interesting orientation for future work. Due to the dynamic patterns of social spammers, it is important to discover attributes separating spammers from non-spammers automatically. We can further investigate methods to improve performance of the proposed model by adding online learning algorithm.

**References:**

1. <https://ggplot2.tidyverse.org/>
2. [**https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city**](https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city)
3. [**https://www.skyfilabs.com/project-ideas/uber-data-analysis**](https://www.skyfilabs.com/project-ideas/uber-data-analysis)

4. Sumner, M., Frank, E., Hall, M.: Speeding up logistic model tree induction. In: Jorge,

A.M., Torgo, L., Brazdil, P.B., Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS (LNAI), vol. 3721, pp. 675–683. Springer, Heidelberg (2005)

5. Yun, Z., Quan, Z., Caixin, S., Shaolan, L., Yuming, L., Yang, S.: RBF neural network and

ANFIS-based short-term load forecasting approach in real-time price environment. IEEE Trans. Power Syst. 23(3), 853–858 (2008)