Spam Detection Framework using ML Algorithm

Abstract- The current use of social media has created incomparable amounts of social data, as it is a cheap and popular information sharing communication platform. Nowadays, a huge percentage of people depend on the accessible material on social networking in their choices (e.g. comments and suggestions about a subject or product). This feature on exchanging knowledge with a wide number of users has quickly prompted social spammers to exploit the network of confidence to distribute spam messages and support personal forums, advertising, phishing, scams and so on. Identifying these spammers and spam material is a hot subject of study, and while large amounts of experiments have recently been conducted to this end, so far the methodologies are only barely able to identify spam feedback, and none of them demonstrates the value of each derived function type. In this study, we have suggested a machine learning-based spam detection system that determines whether or not a specific message in the dataset is spam using a set of machine learning algorithms. Four main features have been used; including user-behavioral, user-linguistic, reviewbehavioral and review-linguistic, to improve the spam detection process and to gather reliable data.

Keywords: Spam Detection, Machine Learning, Random Forest algorithm, Reviews, Framework, Social Media

I. INTRODUCTION

With the advent of technology and everything getting digitalized, we make some of our decisions based on the content of information that we see available on the internet to make the wise or ideal decision to maximize the benefits obtainable when making a choice. From choosing electronic devices to even healthcare products and foods, we tend to check product reviews and pick the one that is most reliable and trustworthy according to the reviews from customers. This in most cases works for the best but there are cases where a fake review or a spam message tends to cheat or divert people away from valid products to potentially harmful or hazardous substances and in some cases even scam gullible people. Spam detection is done manually by designated staffs only when a review is reported as spam by the users of the platform.

This is good in terms of preventing a situation where the system detects the user authentic review in another language as spam and a situation where the system detects the user's authentic review in another language as spam and deletes it instantly but this leads to a lot of potential spam reviews roaming around the site unless it is reported. Some spam reviews are worded right to sound like a normal review but are used as a template to copy-paste everywhere accordingly. This is done clearly to evade from possible reports from other reviewers hence avoiding the possibility of removal completely. Automation of spam detection using a well-defined machine learning framework can greatly help reduce spam reviews that are misleading or fake. Our system uses Machine learning algorithms including Random forest, Bayes Network, Naïve Bayes, K-nearest neighbor and support vector machine combined with NLP techniques to detect and remove spam and to identify the spammer.

II. EXISTING SYSTEM

The current systems of spam detection are solely dependent on three main methods:-

A. Linguistic Based Methods

Humans can comprehend linguistic constructs and their interpretations, but machines can't, and so machines are taught some language in order to help them comprehend linguistic constructs. These techniques are used in search engines to determine the next term in an unfinished sentence. They are split into two Unigrams (Words one by one) and two Bigrams (Words two at a time). As every term has to be remembered, this approach is not as reliable and time intensive.

B. Behavior Based Methods

It is based on Metadata. This method requires users to create a set of laws, and users need to have extensive knowledge of such laws. It needs reformulation because the characteristics of spam shift overtime and the laws need to be modified accordingly. As a consequence, it is mostly user-dependent and still human needs to examine more details.

C. Graph Based Methods

In this approach, by integrating many, heterogeneous details into a single graphical representation, unusual patterns are detected in the data that shows spammer behaviors by running graph-based anomaly detection algorithms for graphical representation. This approach is not reliable, so it is challenging to detect false opinions. Feature engineering is not possible, spam features are not built-in, they are not statistically dependent they are mainly dependent on commercial attractiveness of words and are entirely content-oriented both of these aspects lead to a significant decline of this system.

Spam Detection Framework using ML Algorithm

III. PROPOSED SYSTEM

The system that is proposed on this paper combines random forest algorithm, which is a supervised classification algorithm with NLP concepts to categorize and detect spam reviews among all existing reviews on the TWITTER dataset. There are four major features used in the algorithm which includes 8 NLP concepts:-

A. Review-Behavioral (RB) Based Features This type of functionality is metadata dependent and not the text of the review. There are two aspects to the RB category:-

- **Early Time Frame (ETF)**

Half of the spammers have a very short time span and 55% of the spammers publish all the reviews with a time difference of fewer than 10. That implies the spammers delete their account instantly. Spammers tend to publish their reviews as early as possible, in order to hold their post among the top ratings that many users read first. It can therefore be seen as a guideline for preventing spam.

- **Threshold Rating Deviation**

To determine a reviewer's rating deviation, it measures the total point discrepancy of a company rating point from a consumer ranking. Then we measure the average difference in score for the reviewer in all of his reviews. Spammers also appear to help the firms they have partnered with, so they reward certain organizations with high scores. As a consequence, various companies have a wide variability in their assigned scores which is the reason they have large variation and deviation.

B. Review-Linguistic (RL) Based Features Features in this category are based on the review given by the user and precisely obtained from text. The RL category contains two features:-

- **Ratio of First Personal Pronouns (PP1) and Ratio of Exclamation Sentences (RES)**

Spammers use first personal pronouns and exclamation phrases as much as they can to maximize user's impressions and to emphasize their reviews among others.

C. User-Behavioral (UB) Based Features

Such features are unique to each particular user and are determined by person, meaning that we can use such features to generalize all reviews posted by that same person. This category has two main features:-

- **Burstiness of reviews written by single user**

Spammers usually publish their spam reviews in a limited amount of time for two reasons: one because they intend to influence readers and other people, and the other as they are transient users, they have to write as soon as they can in a limited period of time. A spam may be detected with the aid of the number of comments at the same time. **Average of a user's negative ratio given to different businesses**

Spammers prefer to write reviews that defame firms that compete with those they have partnered with, which may be achieved with negative feedback, or with rating those companies with low scores. Thus, the ratio of their scores appears to be small. This makes it easy to determine whether or not a review is spam.

D. User-Linguistic (UL) Based Features These features are taken from the user's language to demonstrate how customers view their thoughts or views on what they have encountered as a client of a specific company. We use this form of functionality to explain how a spammer interacts in terms of text.

In this category there are two important features:-

- **Average content similitude (ACS) and Maximum content similitude (MCS)**

Spammers usually publish their messages with the same template and tend not to spend their time writing the original review. As a result, they have similar reviews. By contrasting reviews that are similar, a single user can be detected as a bogus user and all of his feedback can be checked and classified as a spam or not.

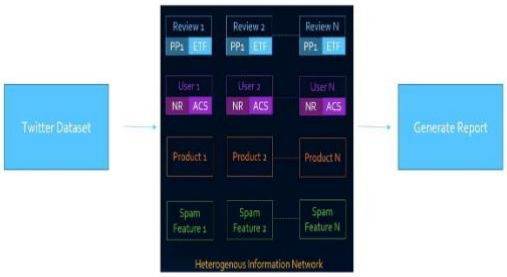

Fig. 1. System Architecture

The proposed framework is introduced with the aid of two key applications, one is anaconda prompt which is exactly similar to the usual command prompt and the other is Jupyter, an integrated python development environment. The anaconda prompt is used for running anaconda and conda commands without changing the directories and to access the local host by connecting the file folder to it and downloading and extracting packages to implement the framework.

Once all packages have been checked and handled, the local host is accessed with jupyter, which includes several code cells.

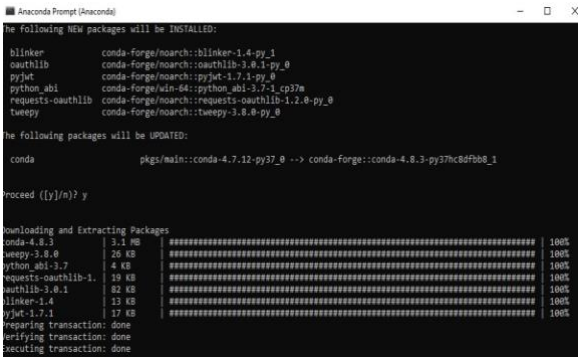

Fig. 2. Anaconda Prompt

These code cells are implemented and run one by one, and the output is obtained after the each cell has been executed.

rating deviation where the diversity and variance of the spammer is checked. Hence, the metadata is collected about the spam

IV. MODULE DESCRIPTION

The proposed framework consists of a set of modules that are implemented:

A. Dataset Extraction

First data is collected from the dataset, in our case which is Twitter messages. After collecting the data, it is cleansed by getting rid of extra spaces, removing duplicates and many more.

B. Collecting Metadata

The RB features are implemented with the cleaned dataset. First, the time frame of the message is identified. After identifying the time frame, it is compared with the threshold

Ivector machine is used.

E. Generating Spam Text Data and information about the Spammer

After the ML algorithms have been implemented the spam messages are identified and obtained, and the information about the spammer who has written the spam message will be collected. With the help of this information, the message and

```
In [24]: print('Name of the Friends of user')
         friends = []
         for friend in tweepy.Cursor(api.friends, screen_name = 'hydcitypolice').items(20):
             try:
                 friends.append(friend.screen_name)
                 print(friend.screen_name)
                 time.sleep()
             except Exception as e:
                 pass

         with open("../Data/Leg_User_txt/friend20.txt", "w") as f:
             for item in friends:
                 f.write("%s\n" % item)
```

```
Name of the Friends of user
TSEduDept
MAHaleem20
spvikarabad
CYBTRAFFIC
ipstelangana
TelanganaCOPs
NORTHRoadSafety
TSPSCofficial
TSMAUDOnline
TSCSOffice
TSConsumers
BlrCityPolice
SpKothaguden
lrvr1974
```

spammer.

C. Generalize Messages

All twitter messages are collected and generalized regardless of whether they are spam or not. By generalizing the messages a lot of time can be saved.

D. Implementing ML Algorithms

The ML algorithms are implemented in this stage by segregating the messages into spam content and original content. ML algorithms including Random forest, Bayes

spammer's entire history can be accessed and all his messages can be analyzed.

V. ADVANTAGES

Feature Engineering is available. Therefore, features of rawdata can be easily extracted with the help of data mining. It is used improve the performance of Machine learning algorithms.

- Each and every data obtained is accurate.
- Spam Features are as a built in function.
- Less human interaction.
- It is statistics based approach.
- Supports review centric spam detection.
- Supports reviewer centric spam detection.

VI. RESULTS AND ANALYSIS

Fig. 4. Spammer Information

The Fig. 4 shows entire data about the spammer including TwitterID, TextData, TweetCreatedAt, RetweetCount, TweetFavouriteCount, TweetSource, UserID, UserScreenName and UserName. This information can inturn help in identifying more spammers with way the text data has been written.

	TwittID	TextData	TweetCreatedAt	RetweetCount	TweetFavouriteCount	TweetSource	UserID	UserScreenName	UserName
11	1239487501414219777	Here's something for all the young ones out th...	2020-03-16 09:43:43	17	46	Twitter for iPhone	1239448814688218560	TSEduDept	Telangana State Education Department
12	1239471638111899648	Happy to note, the majority of the educational...	2020-03-16 08:40:41	12	23	Twitter Web App	1239448814688218560	TSEduDept	Telangana State Education Department
13	1239471401236512769	In view of the #CoronaVirus threat, the Telang...	2020-03-16 08:40:04	16	48	Twitter Web App	1239448814688218560	TSEduDept	Telangana State Education Department
14	1239463906588392448	Greetings!\n\nTelangana State Education Depar...	2020-03-16 08:09:58	82	329	Twitter Web App	1239448814688218560	TSEduDept	Telangana State Education Department

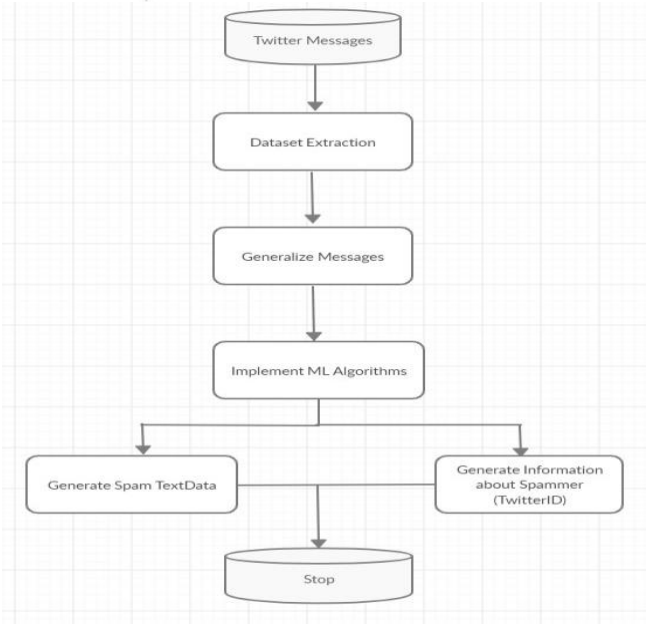

Fig. 5. Names of Spammer's Friends

In Fig.5 the names of the spammer's friends is collected and their profiles are analyzed to make it easier to identify more spammers in future

VII. CONCLUSION

In this paper, we identified the spams and spammers present in a twitter dataset with the help of machine learning algorithms and NLP concepts.

Fig. 3. Flow Diagram

The flow diagram shows the entire flow and steps of the framework.

Spam Detection Framework using ML Algorithm

By reviewing the spam, the entire details about the spammer are accessed and displayed, which in turn helps in determining other spams, spammers and their way of writing messages. We considered two attribute sets which includes content and user behavior, the content is determined with the help of average content similitude, maximum content similitude, ratio of exclamation sentences and the ratio of first personal pronouns. The user behavior is determined with the help of properties such as reviews written and an average of negative ratio given. Thus, making it a very effective and accurate spam detection framework.