EasyChair Preprint

№ 2240

Exploring Feature Selection Technique in Detecting Sybil Accounts in a Social Network

Shradha Sharma and Manu Sood

December 25, 2019

Exploring Feature Selection Technique in Detecting Sybil Accounts in a Social Network

Shradha Sharma[1], Manu Sood[2]
[1]Himachal Pradesh University, Shimla, India
shradha19sharma@gmail.com
[2]Himachal Pradesh University, Shimla, India
soodm_67@yahoo.com

Abstract

The amount of data is increasing rapidly. With the advent in technology social networks are becoming popular day by day. Machine learning provides us the methods to extract useful information. There are different machines learning techniques available. The process of machine learning includes preprocessing, feature selection, building the prediction model and testing the model. In this study we have train a model to detect the Sybil accounts. Since the data is collected from the various resources so preprocessing of the dataset is done in order to remove the noisy data. Feature selection techniques are used for the selection the relevant feature. It removes the redundant and irrelevant features. In this research, we have used three feature selection methods: correlation matrix with heatmap after then feature importance and at last recursive feature elimination with cross validation. Three classifier were used to train the model. Those are random forest, support vector machine and k nearest neighbour. We have used different metrics to evaluate the results obtained from classifier. In our study we conclude that the Random Forest provides the best results out of three classifier which have been used.
Keyword used : Preprocessing, Feature Selection, Classification, K Nearest Neighbour, Random Forest

1. Introduction

With the passage of time the amount of data is increasing. It seems very difficult to extract some useful information from such huge amount of data. But Machine Learning provides us the way with the help of which we can extract useful information from a set of large amount of data. Machine Learning is the subset of the Artificial intelligence. Machine learning focuses on design of an algorithm which learns from the provided data.

There are four types of machine learning techniques named as: Supervised learning, Unsupervised learning, Semi-supervised learning and Reinforcement learning. Supervised learning train the machine by using labelled data. It mainly uses classification and regression techniques. Classification predict the discrete responses. Regression predict the continuous responses. Unsupervised learning do not consists labelled data. It uses techniques like clustering and association. Clustering is mainly used for the grouping of data and finding hidden patterns. Association rules mainly helps in finding out the association among the data objects in huge dataset. In Semi-supervised learning the amount of input data is very large but only small data is labeled. It is used with methods classification, regression and prediction. It mainly used graph based methods. Reinforcement learning mainly based on the reward and punishment method. It consists of three main components: agent, environment and actions. It is mainly used in gaming, navigation and robotics.

The process of training a model with the help of machine learning consists of various step. The very first step is collection of data which is necessary to train the model. Data is collected from various sources. Since the data collected from various sources is large in size and may consists of noise. So next step is preprocessing of the data. Preprocessing of data mainly consists of data cleaning, data transformation, error correction and normalization [1]. Data cleaning deals with the removal of missing and noisy data. Missing data is the data in which some values are missing in the data and noisy data is irrelevant data.
The next one is data transformation which transforms data into the required form.

The next step is feature selection. Dataset may consists of hundreds of features. Out of which only few are of use. So in feature selection technique irrelevant features are dropped. This mainly improves the accuracy and reduces the training time. There are three methods for the feature selection: Filter technique, Wrapper technique and Embedded Technique.
In filter selection features are selected on the basis of their scores in statistical test for their correlation with target variable [W1]. In wrapper method feature subset is selected by using induction algorithm. In wrapper method induction algorithm is also the part of evaluation function [2]. The third technique is embedded method. This method combines the filter and wrapper method.

Next step in the modeling is selection of ML technique. There are number of techniques available for training the model. Since we have labeled data and our problem is related to the classification model so we have used Random Forest, Support Vector Machine and K-Nearest Neighbor prediction model. After the selection of prediction model the next important step is to test the prediction model. In order to do that dataset is divided into three parts: training data and testing data. Training data is used to train the model and testing data is used to test the designed model.

Social network is a very huge platform which provides the way by which people all around the world can connect with each other. Online social network (OSN) provides a platform through which people can exchange their ideas and information with people all around world. Online social network are the type of virtual communities. These virtual communities are connected through hardware, networking and special software for socialization. The most popular OSN used by people is twitter.
The OSNs either provides some rules to detect the spambots and to remove it, or it provides some software program for same purpose, otherwise if it is fail to detect the spambot then spambot is successful in their malicious design. [3, 4, 5]

Twitter, originally started as the personal microblogging site [6].With the advancement in technology the number of users start increasing. But in order to increase the number of followers of the target account fake followers and social spambots comes into the play. On the other hand social spambot is a profile on the social media platform that is automatically programmed to generate messages, follow accounts.

In this paper authors have explored the mechanism to identify weather the given account is human account, fake account or spambot.

The goal of this study are as following:

1) **To preprocess the dataset, in order to select best optimal features from the feature set.**
2) **To compare the performance the two machine learning classifier i.e. RF,KNN**
3) **To study and analyze the effect of biasing on human account with fake followers and human account with social spambots**
4) **To study the effect of overfitting on biased which is data under consideration.**

Novelty of the work: The dataset contains both numerical and categorical values. Each and every instance in the dataset have unique value .Since we have used python for the implementation purpose it could not deal with the categorical values. Instead of assigning binary values to the categorical data, all the categorical values are converted into its corresponding numerical value, which in return provides the better results.

In the process of biasing, human accounts, fake accounts and spambots are varied proportionally to predict each and every case.
For the evaluation purpose 6 matrices are used.

1.1 Roadmap

The remainder of this paper consists of following sections. Section 2 provides the information about the related work. Section 3 consists of the brief discussion about the datasets used by authors in their research process. Section 4 is about the methodology. Section 5 provides the

results .Section 6 provides the conclusion and future scope and last section provides the future scope.

2. Related Work

Machine learning is being widely used across the world in different problem domain. The techniques of machine learning have also been used in predicting the fake twitter accounts and spambots. Some of related work has been listed down so as to get an idea about the work that has been carried out.

For the detection of fake twitter account a study have been carried out. In this study authors have used 49 distinct features. 8machine learning classifiers were used to conduct the study. [6].

Work in [7], they measure how much a current twitter is capable to detect social spambots. Later, they measure the human performance for detecting difference between genuine account, social spam bots, and traditional spam bots.

In an another study 14 features has been selected by authors by using recursive feature elimination technique for feature selection and then uses machine learning classifiers for the detection of fake twitter account [8].

In [6], the authors list some criteria with the help of which one can detect clients and victims of Twitter account market.

3. Experiment setup

For the implementation of various machine learning algorithms python language was used. We have used Jupyter Notebook to execute the codes written by using python language. Jupyter Notebook is web based application and it is open source. [W2].

3.1 Dataset

In order to achieve the number of objectives above listed, a number a datasets containing human accounts, fake twitter accounts and spambots have used. Total nine datasets used in this study, out of which three datasets have listed human accounts, three datasets have listed the fake accounts data and other three contain the data for spambots. The feature set in each dataset have same labels and same number of features. Table 1 provides the description about the type of dataset, nature of dataset and number of accounts in each dataset.

Table1: Description of Datasets

S.No	Dataset	Nature of accounts dataset contains	No. of accounts
1	E13 (elezioni2013)	Human	1481
2	TFP (TheFakeProject)	Human	469
3	Genuine Accounts	Human	3474
4	TWT (twittertechnology)	Fake accounts	845
5	INT (intertwitter)	Fake accounts	1337
6	FSF (fastfollowerz)	Fake accounts	1169
7	Spambot 1	Spambots	991
8	Spambot 2	Spambots	3457
9	Spambot 3	Spambots	464

Table 2: Description of Featureset

Id	Friends count	Language	Geo enabled
Name	Favorite count	Time zone	Profile image url
Screen name	Listed count	Location	Profile banner url
Status count	Created at	Default profile	Profile text color
Followers count	url	Default profile image	Profile image url https
Utc offset	Protected	Verified	Updated
Profile sidebar fill color	Profile background image url	Profile background color	Profile link color
Profile use background image	Profile background image url https	Profile sidebar border color	Profile background

The authors of [6] have created this dataset for their study. They have verified each and every genuine account themselves. The first dataset of human account i.e. E13 (elezioni2013) has been created during the elections conducted in Italy. The second dataset of human account i.e. TFP (The Fake Project) is created by authors themselves. They had stared a project named “The Fake Project” (a twitter account) in order to collect the human account. The next three datasets which contains the fake account details were bought online by the same authors.

3.2 Data Preprocessing

Data pre-processing is the process of cleaning, scaling and transforming the data into required format. In the process of data cleaning NaN (Not-a-Number), inconsistent and missing values are removed [1].

The data which is under consideration during this whole work also had some missing values and there are also some features which didn't consists any value. The first step was removal of those feature which didn't contain any values. After that missing values were replaced with zero. Features contained the redundant values were also dropped. After the data preprocessing we got 24 features into the feature set.

Now that selected feature subset was further processed by using feature selection techniques.

3.3 Feature Selection

The main reason behind to take large dataset is that, it contains large number of features in the feature set. These features contained the information about the target variables. There is a myth if you have more number of features in feature set then you will get better performance. But this is not valid for every case. In feature set there are some features, if we remove them from the feature set, they don't create any problem to our solution. They are generally irrelevant, noisy, and redundant features [10].
There are three feature selection methods: Filter method, Wrapper method, Embedded method. Filter method mainly uses the mathematical functions. It is faster than wrapper method. It mainly includes univariate method, chi square method and correlation matrix with heatmap. Wrapper methods, uses the classifier to prepare the feature set with maximum accuracy. The accuracy provided by wrapper methods is better [8]. Third method is embedded method which consists the quality of other both method. Filter method provide better results if there are very large number of features in the feature set. But if we need to deal with fewer features then, wrapper methods work better [8].

Three feature selection techniques have been used in order to find out the best optimal features. The first one is correlation matrix with heatmap. Correlation matrix mainly defines the relationship between the features themselves and with target variable. The second method used is feature importance. This method ranks the features according to their importance.The third method used is recursive feature elimination with cross validation (rfe_cv).
RFE with cross validation works approximately this way. You have to specify as input: a classifier C you want to use for prediction (e.g., a decision tree), a scoring function F to evaluate the performance of the classifier (e.g., accuracy); how many features k you want to eliminate in every step (e.g. 1 feature) [W3]. Then RFE with cross validation starts with all the n features, makes predictions with cross validation using C, computes the relative cross-validated performance score F and the ranking of the importance of the features. Then it eliminates the

lowest k features in the ranking and re-makes the predictions, the computation of the performance score and the feature ranking. It proceeds until all the features are eliminated. Finally it outputs the set of features which produced the predictor with the best performance score [W3].

After applying correlation matrix with heatmap 15 features were selected out of 24 features. After that second feature selection technique is applied. From second feature selection technique i.e RFE_CV technique 8 best optimal features were selected. All the operations are performed on these selected best optimal features.

3.4 Dataset Biasing

Table 3: Biased Datasets

S.no	Case	Human accounts	Fake accounts	Spam bot	Total
1	(E13-FSF) (50%-50%)	741	585	-	1326
2	(E13-FSF) (25%-75%)	371	877	-	1248
3	(E13-FSF) (75%-25%)	1111	293	-	1404
4	(E13-INT) (50%-50%)	741	669	-	1410
5	(E13-INT) (25%-75%)	371	1003	-	1374
6	(E13-INT) (75%-25%)	1111	335	-	1446
7	(E13-TWT) (50%-50%)	741	423	-	1164
8	(E13-TWT) (25%-75%)	371	634	-	1005
9	(E13-TWT) (75%-25%)	1111	212	-	1323
10	(Genuine-spambot1) (50%-50%)	1738	-	496	2234
11	(Genuine-spambot1) (25%-75%)	869	-	744	1613
12	(Genuine-spambot1) (75%-25%)	2609	-	248	2857
13	(Genuine-spambot2) (50%-50%)	1738	-	1729	3467
14	(Genuine-spambot2) (25%-75%)	869	-	2594	3469
15	(Genuine-spambot2) (75%-25%)	2609	-	865	3474
16	(Genuine-spambot3) (50%-50%)	1738	-	233	1971
17	(Genuine-spambot3) (25%-75%)	869	-	349	1218
18	(Genuine-spambot3) (75%-25%)	2609	-	117	2726

The biasing of data is carried out as follows:

1) 50% human account and 50% fake accounts.
2) 25% human account and 75% fake accounts.
3) 75% human account and 25% fake accounts.
4) 50% human account and 50% spambots.

5) 25% human account and 75% spambots.
6) 75% human account and 25% spambots.

Table 3 consists the complete description about the biased dataset.

3.5 Classifier used

Three classifier were used for the study the first one is Random Forest, Support Vector Machine and K- Nearest Neighbour.

Random forest is used to solve both the problems i.e. classification and regression. This algorithm works in two steps, in first step random tress are created, and the second step helps in the prediction from the votes of trees created in the first step to determine the output [1].

KNN is also known as non-parametric algorithm. It is also used for both classification and regression. The KNN works on the principle that the objects within a dataset are close to each other and have similar properties [11]. KNN is also used for both classification and regression. This algorithm is very faster to train. For the purpose of training and testing the models dataset is used in the ratio of 70/30. 70% of the dataset is used for training the model and 30% of the dataset is used for testing purpose.

SVM (Support Vector Machine) is a discriminative classifier which is used for the classification and regression in machine learning algorithm. It constructs the hyperplane between the data objects. There can be multiple boundaries which can separate the data points but best out of all the boundaries is selected. The selected best boundary is known as hyperplane. Hyperplane is selected in such a manner that there should be maximum distance between the data points. Data points close to hyperplane and effects its position are known as support vector.

3.6 Evaluation criteria

The final outcomes of the experiments were evaluated on the basic of some metrics that uses the following indicators [12]:
True Positive (TP). It indicates the number of fake followers those are identified as fake.
True Negative (TN). It indicates the number of human followers those are identified as human.
False Positive (FP). This indicates the number of human followers those are identified as fake.
False Negative (FN). This indicates the number of fake followers those are identified as human.
The matrix formed with the help of these metrics is called confusion matrix. This matrix compares the actual class with the predicted class which can be positive and negative.

Evaluation metrics:

In order to evaluate the final result following metrics were used [12]:

- **Accuracy**: Ratio of predicted true results in the selected dataset, calculated as:

(TP+ TN) ÷ (TP+TN+FP+FN).

- **Precision:** Ratio of identified positive cases which were indeed positive, calculated as: TP ÷ (TP+FP).

- **Recall:** Ratio of real positive cases that are indeed identified positive, calculated as: TP ÷ (TP+FN).

- **F-Measure**: It is harmonic mean of recall and precision, calculated as: (2*precision*recall)/ (precision + recall).

- **Matthew Correlation Coefficient (MCC):** Estimates the correlation between the identified class and the real class of samples, calculated as: (TP*TN-FP*FN) ÷ √((TP+FN)(TP+FP)(TN+FP)(TN+FN)).

- **Specificity**: It is the ratio of identified true negative and sum of true negative and false positive, calculated as: TN/ (TN+FP).

Overfitting is a situation in which a training dataset is trained in such a way that it provides good accuracy but in case of testing dataset the accuracy is not so good.

5 Results and analysis

After selecting the features by using feature selection techniques, the selected features with their data were used for the experiment. Prediction model was trained with three classifiers: Random forest, KNN and Support Vector Machine. The prediction model was tested on 18 datasets. The results collected from the experiment are listed in tables 4, 5 and 6. Table have complete description about the computed evaluation metrics. Table 4 shows the performance of predictive model trained by using KNN. Table 5 shows the performance of predictive model trained by using RF. Table 6 shows the performance of predictive model trained by using SVM.

The highest value for every metric is shown in blue box and lowest value for metric is shown in green box. In case of KNN, we got best result for dataset (E13-FSF) (75%-25%).

In case of random forest the value for each metric is achieved 1.00 except precision (0.99) for dataset (E13-FSF) (50%-50%).

Table 4: Evaluation Metrics for KNN

Dataset Case	K Nearest Neighbour									
	Confusion matrix				Evaluation metrics					
	TN	FP	FN	TP	Accuracy	Precision	Recall	F-Measure	MCC	Specificity
(E13-FSF) (50%-50%)	163	0	1	234	0.99	1	0.99	0.99	0.99	1
(E13-FSF) (25%-75%)	288	0	1	107	0.99	1	0.99	0.99	0.98	1
(E13-FSF) (75%-25%)	95	0	3	324	0.99	1	1	0.99	0.99	1
(E13-INT) (50%-50%)	199	0	16	767	0.98	1	1	0.97	0.97	1
(E13-INT) (25%-75%)	301	4	6	99	0.97	0.96	0.94	0.93	0.96	0.98
(E13-INT) (75%-25%)	299	7	12	98	0.96	0.93	0.89	0.92	0.93	0.97
(E13-TWT) (50%-50%)	175	2	4	110	0.97	0.98	0.96	0.96	0.96	0.98
(E13-TWT) (25%-75%)	234	0	1	102	0.99	1	0.99	0.99	0.99	1
(E13-TWT) (75%-25%)	281	2	7	86	0.97	0.97	0.92	0.94	0.95	0.99
(Genuine-spambot1) (50%-50%)	152	0	3	112	0.98	1	0.97	1	0.97	1
(Genuine-spambot1) (25%-75%)	241	0	2	211	0.99	1	0.99	0.99	0.98	1
(Genuine-spambot1) (75%-25%)	189	1	4	187	0.98	0.99	0.97	0.97	0.97	0.99
(Genuine-spambot2) (50%-50%)	231	2	3	97	0.98	0.97	0.97	0.94	0.96	0.99
(Genuine-spambot2) (25%-75%)	271	4	2	121	0.98	0.97	0.98	0.97	0.98	0.97
(Genuine-spambot2) (75%-25%)	177	3	1	217	0.99	0.98	0.99	0.98	0.98	0.98

Table 5: Evaluation Metrics for Random Forest

Dataset Case	Random forest									
	Confusion matrix				Evaluation metrics					
	TP	FN	FP	TN	Accuracy	Precision	Recall	F-Measure	MCC	Specificity
(E13-FSF) (50%-50%)	189	0	1	247	1	0.99	1	1	1	1
(E13-FSF) (25%-75%)	266	0	0	126	1	1	1	1	1	1
(E13-FSF) (75%-25%)	94	0	0	370	1	1	1	1	1	1
(E13-INT) (50%-50%)	222	0	0	245	1	1	1	1	1	1
(E13-INT) (25%-75%)	332	0	0	122	1	1	1	1	1	1
(E13-INT) (75%-25%)	241	0	0	770	1	1	1	1	1	1
(E13-TWT) (50%-50%)	169	0	0	231	1	1	1	1	1	1
(E13-TWT) (25%-75%)	240	0	0	111	1	1	1	1	1	1
(E13-TWT) (75%-25%)	88	0	0	343	1	1	1	1	1	1
(Genuine-spambot1) (50%-50%)	123	0	0	123	1	1	1	1	1	1
(Genuine-spambot1) (25%-75%)	342	0	0	145	1	1	1	1	1	1
(Genuine-spambot1) (75%-25%)	176	0	0	76	1	1	1	1	1	1
(Genuine-spambot2) (50%-50%)	145	0	0	87	1	1	1	1	1	1
(Genuine-spambot2) (25%-75%)	242	0	0	132	1	1	1	1	1	1
(Genuine-spambot2) (75%-25%)	139	0	0	45	1	1	1	1	1	1
(Genuine-spambot3) (50%-50%)	217	0	0	134	1	1	1	1	1	1
(Genuine-spambot3) (25%-75%)	132	0	0	199	1	1	1	1	1	1
(Genuine-spambot3) (75%-25%)	169	0	0	156	1	1	1	1	1	1

In case of SVM, from fourteen datasets we got best results and only four dataset have provided values less than others, and those are (E13-FSF) (75%-25%),(E13-TWT) (50%-50%),(Genuine-spambot1) (75%-25%),(Genuine-spambot3) (50%-50%).

After comparing maximum values obtained from different three classifier we observed that Radom forest is best amongst all the three classifier and we got best results for dataset (E13-FSF) (75%-25%) in all three cases.

Table 6: Evaluation Metrics for SVM

Dataset Case	SVM									
	Confusion matrix				Evaluation metrics					
	TN	FP	FN	TP	Accuracy	Precision	Recall	F-Measure	MCC	specificity
(E13-FSF) (50%-50%)	156	0	1	227	0.99	1	0.99	0.99	0.99	1
(E13-FSF) (25%-75%)	281	0	1	100	0.99	1	0.99	0.99	0.99	1
(E13-FSF) (75%-25%)	192	0	2	260	0.99	1	0.99	0.99	0.99	1
(E13-INT) (50%-50%)	88	2	5	317	0.98	0.99	0.99	0.98	0.98	0.97
(E13-INT) (25%-75%)	294	0	1	92	0.99	1	0.99	0.99	0.99	1
(E13-INT) (75%-25%)	292	0	1	91	0.99	1	0.99	0.99	0.99	1
(E13-TWT) (50%-50%)	168	1	4	103	0.98	0.99	0.96	0.97	0.97	0.99
(E13-TWT) (25%-75%)	227	0	1	95	0.99	1	0.99	0.99	0.99	1
(E13-TWT) (75%-25%)	274	0	1	79	0.99	1	0.99	0.99	0.99	1
(Genuine-spambot1) (50%-50%)	231	0	1	89	0.99	1	0.99	0.99	0.99	1
(Genuine-spambot1) (25%-75%)	143	0	1	134	0.99	1	0.99	0.99	0.99	1
(Genuine-spambot1) (75%-25%)	176	2	4	123	0.98	0.98	0.96	0.96	0.97	0.98
(Genuine-spambot2) (50%-50%)	185	0	1	111	0.99	1	0.99	0.99	0.99	1
(Genuine-spambot2) (25%-75%)	243	0	1	165	0.99	1	0.99	0.99	0.99	1
(Genuine-spambot2) (75%-25%)	159	0	1	67	0.99	1	0.99	0.99	0.99	1
(Genuine-spambot3) (50%-50%)	189	2	4	121	0.98	0.98	0.96	0.96	0.97	0.98
(Genuine-spambot3) (25%-75%)	231	0	1	189	0.99	1	0.99	0.99	0.99	1
(Genuine-spambot3) (75%-25%)	201	0	1	154	0.99	1	0.99	0.99	0.99	1

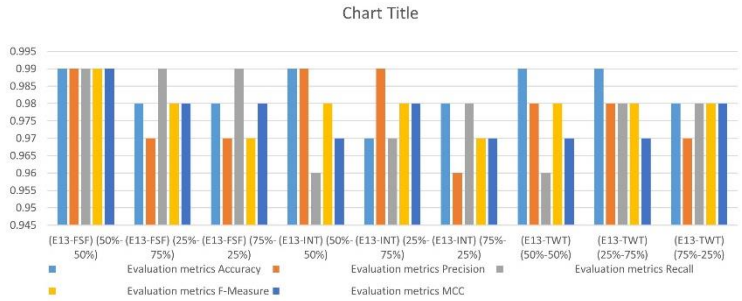

Figure 1: Bar Graph for Evaluation Metrics for SVM

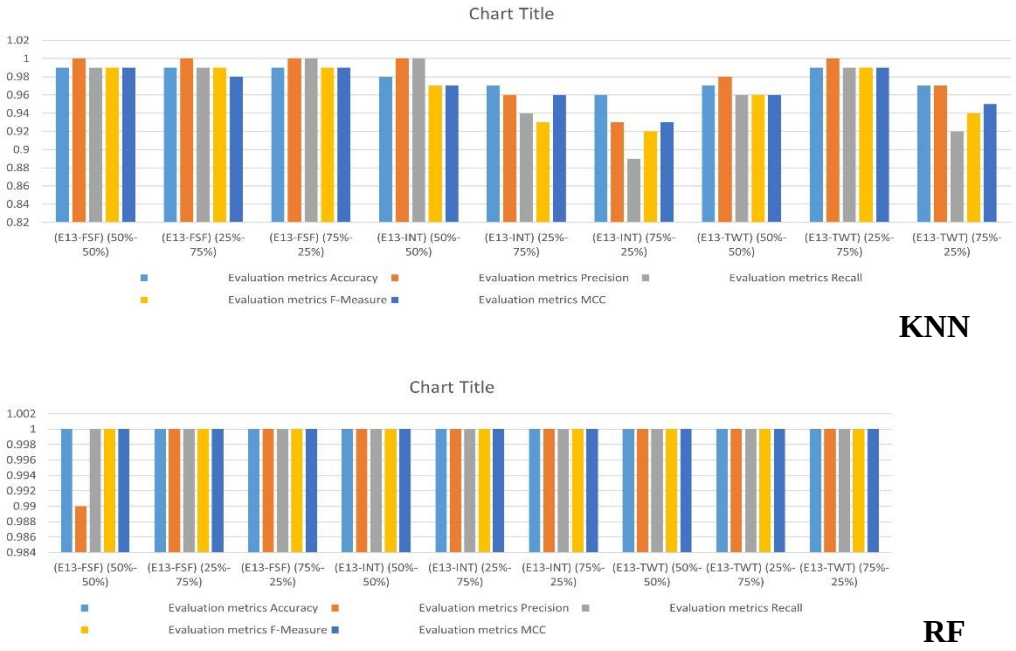

Figure 2: Bar Graph of Evaluation Metrics for KNN and RF

6. Future scope and conclusion

In this work we have firstly used preprocessing on the data collected from various sources and after that feature selection techniques were used on preprocessed data **.** Preprocessing removes all the noisy data and Feature selection removes all the redundant features and helps in selecting the best optimal features. After the selection of best optimal features a predicted model is created with the help of three different classifier. Out of three different classifier used, RF provides the best results. We can further extend this study to implement high performance model for real time environment. We can use this study to design some set of rules which can help us to detect fake human accounts and spambots.

7. Acknowledgement

Our utmost gratitude towards the Cresci et al. [6] to allow us use the dataset created by them. Since this dataset was the basic requirement for this research. They allowed us to use their dataset in our study.

8. References

1. N. Bindra and M. Sood (2018), “Data pre-processing techniques for boosting performance in network traffic classification”, First International Conference on Computational

Intelligence and Data Analytics, ICCIDA-2018 26-27 October 2018, Springer CCIS Series, Gandhi Institute For Technology (GIFT), Bhubaneshwar, Odhisha, India.

2. G. John, R. Kohavi and K. Pfleger, Irrelevant features and the subset selection problem, in: Proceedings Fifth International Conference on Machine Learning, New Brunswick, NJ (Morgan Kaufmann, Los Altos, CA, 1994) 121-129.

3.Vasudeva, A., Sood, M.: Survey on Sybil attack defense mechanisms in wireless ad hoc networks. Journal of Network and Computer Applications (2018)

4. Vasudeva, A., Sood, M.: A Vampire Act of Sybil Attack on the Highest Node Degree Clustering in Mobile Ad Hoc Networks. Indian journal of science and technology, vol 9 (2016) .

5. Vasudeva, A., Sood, M.: Perspectives of Sybil attack in routing protocols of mobile ad hoc network. Computer Networks & Communications (NetCom). Springer, New York, NY, pp. 3-13 (2013)

6. S. Cresci, R. D. Pietro, R. Petrocchi, A. Spognardi, and M. Tesconi (2015), "Fame for sale: Efficient detection of fake Twitter followers," Decision Support Systems, 80, pp. 56-71.

7. S. Cresci, R. D. Pietro, R. Petrocchi, A. Spognardi, and M. Tesconi (2017), "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," In Proceedings of the 26th International Conference on World Wide Web Companion International World Wide Web Conferences Steering Committee, pp. 963-972.

8. D. Sonkhla and M. Sood(2019), "Performance Analysis and Feature Selection on Sybil User Data using Recursive Feature Elimination ", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-9S4, July 2019, pp. 48-56.

9. G. Stringhini, M. Egele, C. Kruegel, G. Vigna, Poultry markets: on the underground economy of Twitter followers, Workshop on Online Social Networks, WOSN'12, ACM 2012, pp. 1–6.

10. K. Yan, and D. Zhang (2015), "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," Sensors and Actuators B: Chemical, 212, pp. 353-363.

11. S. B. Kotsiantis, I. Zaharakis, and P. Pintelas (2007), "Supervised machine learning: A review of classification techniques," Emerging artificial intelligence applications in computer engineering, 160, pp. 3-24.

12. M. Alsaleh, A. Alarifi, A. M. Al-Salman, M. Alfayez, and A. Almuhaysin (2014), "Tsd: Detecting sybil accounts in twitter," In 2014 13th International Conference on Machine Learning and Applications, IEEE, pp. 463-469.

Web references

W1. https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/accessed on 21/11/2019/9:30pm IST.

W2. Project Jupyter. Available: https://jupyter.org/. Last Accessed on 21/11/2019.

W3. https://www.researchgate.net/post/Recursive_feature_selection_with_cross-validation_in_the_caret_package_R_how_is_the_final_best_feature_set_selected2/ accessed on 22/11/ 2019