A Project report on

Detection of Fake and Clone Accounts in Twitter using Classification and Distance Measure Algorithms

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

G. KIRAN DEEPAK
(19H51A0570)

K.V.S.S.S.R.H SRI HARSHA
(19H51A0573)

G. PRANAY KUMAR
(19H51A05F8)

Under the esteemed guidance of

A. POONGODAI
(Associate Professor)

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

(An Autonomous Institution under UGC & JNTUH, Approved by AICTE, Permanently Affiliated to JNTUH, Accredited by NBA.)

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2019- 2023

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Major Project Phase-1 report entitled **"Detection of fake and clone accounts using classification and distance measure algorithms"** being submitted by G,Kiran Deepak (19H51A0570), K.V.S.S.S.R.H Sri Harsha (19H51A0573), G.Pranay Kumar (19H51A05F8) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Dr.A.Poongodai
Associate Professor
Dept. of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. of CSE

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Dr.A.Poongodai, Associate Professor** , Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academic, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the Teaching & Non- teaching staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Mr. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

G.Kiran Deepak	19H51A0570
K.V.S.S.S.R.H Sri Harsha	19H51A0573
G.Pranay Kumar	19H51A05F8

TABLE OF CONTENTS

List of Figures

List of Tables

ABSTRACT

Online Social Network (OSN) is a network hub where people with similar interests or real world relationships interact. As the popularity of OSN is increasing, the security and privacy issues related to it are also rising. Fake and Clone profiles are creating dangerous security problems for social network users. Cloning of user profiles is one serious threat, where already existing user's details are stolen to create duplicate profiles and then it is misused for damaging the identity of the original profile owner. They can even launch threats like phishing, stalking, spamming etc. For Profile Cloning detection two methods are used. One using Similarity Measures and the other using C4.5 decision tree algorithm. In Similarity Measures, two types of similarities are considered – Similarity of Attributes and Similarity of Network relationships. C4.5 detects clones by building decision tree by taking information gain into consideration. A comparison is made to check how well these two methods help in detecting clone profiles.

CHAPTER 1
INTRODUCTION

CHAPTER 1

INTRODUCTION

ONLINE Social Networks (OSN) like Face book, Twitter, LinkedIn, Instagram etc are used by billions of users all around the world to build network connections. The ease and accessibility of social networks have created a new era of networking. OSN users share a lot of information in the network like photos, videos, school name, college name, phone numbers, email address, home address, family relations, bank details, career details etc. This information if put into hands of attackers, the after effects are very severe.

Most of the OSN users are unaware of the security threats that exist in the social networks and easily fall prey to these attacks. The risks are more dangerous if the victims are children. In Profile Cloning attack, the profile information of existing users are stolen to create duplicate profiles and these profiles are misused for spoiling the identity of original profile owners[1- 6]. There are two types of Profile Cloning namely - Same Site and Cross Site Profile Cloning[1,7-9]. If user credentials are taken from one Network to create a clone profile in same Network then it is called Same Site profile cloning[1,10-12].

In Cross Site profile cloning, attacker takes the user information from one Network to create a duplicate profile in other Network in which the user is not having any account[1,13-15]. As the registration process in social networks have become very simple in order to attract more and more users, the creation of fake profiles are also increasing in an alarming rate. An attacker creates a fake profile in order to connect to a victim to cause malicious activities. And also to spread fake news and spam messages.

Problem Statement

Today, Fake and Clone profiles have become a very serious threat in social networks. So, a detection method is very much necessary to find these frauds who use people's faith to gather private information and create duplicate profiles. Many authors have worked in this area and have proposed methods to identify these type of profiles in social networks.

Research Objective:

The main objective of this project is to

1. Generate an algorithm to detect fake accounts from the set of data given to the classifier

2. Generate an algorithm to detect clone accounts from the set of data given to the classifier.

3. Generate an efficient output than the existing solutions present in the market.

Project Scope And Limitations:

Project Scope

- Our Problem falls under classification category.
- In our project, features in the database created for review websites are classified by determining the input and output parameters for the classification.
- Classification is to determine the class to which each data sample of the methods belongs, which methods are used when the outputs of input data are qualitative.
- For the present the project is limited to twitter accounts data set which classifies the data of twitter segregates and does the analysis on twitter data set.

Limitations

- Our Project can achieve an accuracy of 90%.
- Our Project can classify on data of only certain features.
- Our project is only used to detect the fake and clone accounts of only the twitter data set.

CHAPTER 2
BACKGROUND WORK

BACKGROUND WORK

2.1 Fake Account Detection using Naïve Bayes Algorithm

2.1.1 Introduction

Naive Bayes Bernoulli is a binary independence model, which generates an indicator for each term of the vocabulary, either 1 indicating presence of the term in the document or 0 indicating absence. Bernoulli model uses binary occurrence information, ignoring the number of occurrences whereas the multinomial model keeps track of multiple occurrences. It specifies that a review is represented by a vector of binary attributes indicating which accounts are in the fake classification or not.

2.1.2 Merits, Demerits and Challenges

Merits

- It is easy and fast to predict class of test data set. It also performs well in multi class prediction.
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.

Demerits

- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”.
- Accuracy of Naïve bayes model is 75%.

Challenges

Accounts dataset may contain large number of features and the model will not function as expected.

2.1.3. Implementation of Naïve bayes model

- Assume that all features in data set are independent
- Choose the most appropriate dependent variable and split the data set into training, validation, and testing.
- Build a frequency count of the training set and apply Naïve bayes on it

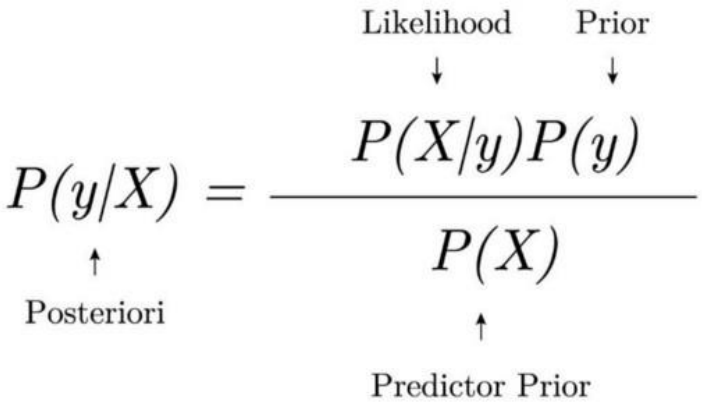

Figure 2.1.3.1: Naïve Bayes

The Bayes' theorem is used to determine the probability of a hypothesis when prior knowledge is available. It depends on conditional probabilities. The formula is given below:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where P(A|B) is posterior probability i.e. the probability of a hypothesis A given the event B occurs. P(B|A) is likelihood probability i.e. the probability of the evidence given that hypothesis A is true. P(A) is prior probability i.e. the probability of the hypothesis before observing the evidence and P(B) is marginal probability i.e. the probability of the evidence. When the Bayes' theorem is applied to classify accounts, the class c of a particular account d is given by :

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Dropping the denominator

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

Document d represented as features x1..xn

2.2 Fake Account Detection using SVM

2.2.1 Introduction

"Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).

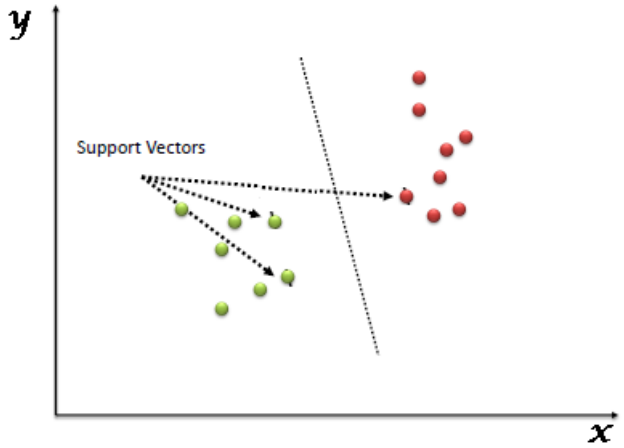

Figure 2.2.1.1: Graph of hyper plane

Support Vectors are simply the coordinates of individual observation. The SVM classifier is a frontier that best segregates the two classes (hyper-plane/ line).

2.2.2 Merits, Demerits, And Challenges

Merits

- It works well with a clear margin of separation
- It is effective in high-dimensional spaces.
- It is effective in cases where the number of dimensions is greater than the number of samples.
- It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

Demerits

- It doesn't perform well when we have large data set because the required training time is higher
- It also doesn't perform very well, when the data set has more noise i.e. target classes are overlapping
- SVM doesn't directly provide probability estimates, these are calculated using an expensive five-fold cross-validation. It is included in the related SVC method of the Python scikit-learn library.

Challenges

- Unbalanced Data
- Multi-label classifications
- SVM cannot handle large data sets
- Semi-supervised learning

2.2.3 Implementation of SVM Model

- Extract the feature vector. Convert text to vector matrix
- Train SVMs based on the saved sample database.
- Analyze the Accounts features by the set of SVMs trained in advance.
- If there are no more unclassified samples, then STOP. Otherwise, continue model training
- Add these test samples into their corresponding database for further training.

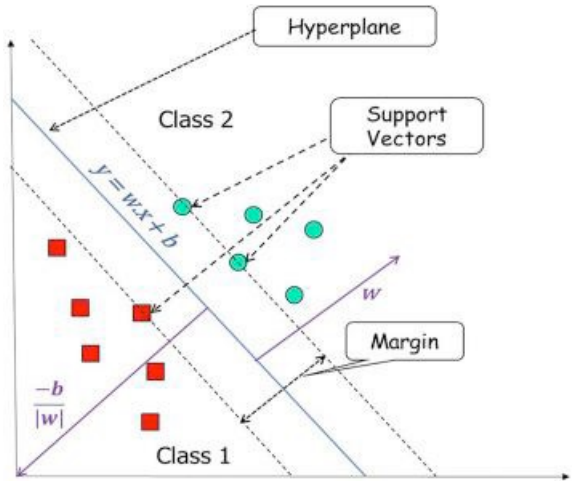

Figure 2.2.3.1: SVM algorithm

2.3 Fake Account detection using Random Forest Algorithm

2.3.1 Introduction

Random Forest is a machine learning algorithm that considered as a supervised learning technique. It creates several Decision Trees on the subset of data. As well as it combines each tree prediction to give the final output prediction for the whole tree based on all votes technique (Jehad, Rehanullah, Nasir and Imran, 2012) (Pretorius, Bierman and Steel, 2016). Moreover, Random Forest is used in Regression and Classification of ML.

2.3.2Merits, Demerits and Challenges

Merits

- It is proved the effectiveness of this algorithm on large datasets compared to other classifiers like: Neural Networks, Discriminant
- Analysis and Support Vector Machines (SVM) (Jehad, Rehanullah, Nasir and Imran, 2012).
- One of the most important benefits of Random Forest is that it can work with missing data, which is the replacement of missing values by the variable that is common in a particular node.
- The Random Forest can also handle big data quickly, provide a higher accuracy and prevent overfitting problems.

Demerits

- Random Forest requires many computational resources and large memory for storage, due to the fact that it creates a lot of trees to save information piped generated from hundreds of individual trees.

Challenges

- Larger Computation time as a larger tree need to be constructed and many decisions need to be taken
- Large storage or memory required for storing large tree.

2.3.2 Implementation of Random Forest Algorithm

While building a Random Forest tree classifier, the main thing is to select the best attribute from the total features list of the dataset for the root nodes as well as for sub-nodes. The selection of best attributes is being achieved with the help of a technique known as the Attribute selection measure (ASM).

- Select the best Features using Attribute Selection Measures (ASM) to split the records.
- Make that attribute/feature a decision node and break the dataset into smaller subsets.
- Start the tree-building process by repeating this process recursively for each child until one of the following condition is being achieved:

  a) All tuples belonging to the same attribute value.

  b) There are no more of the attributes remaining.

  c) There are no more instances remaining.

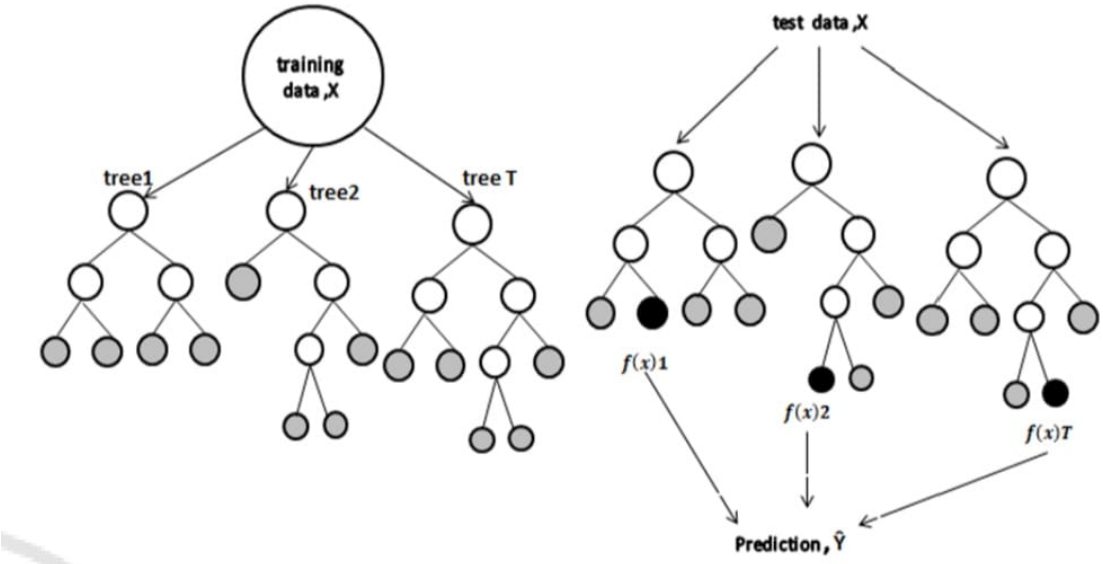

Figure 2.3.3.1 Architecture of Random Forest Classifier

CHAPTER 3
RESULTS AND DISCUSSION

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Comparison of existing Solution

From the existing solution we can see that random forest has a higher accuracy among all the existing solutions but the tree constructed is larger which results in delay in the processing and more consumption of memory. Next for the Naïve bayes algorithm is that it supposes that all features are separated or not connected. So, we cannot know the relationship between features. This algorithm faces a challenge ‘zero-frequency trouble’, (i.e. if the categorical variable has a category in the testing dataset, but not observed in the training dataset, then the pattern assigns a 0 probability and will not able to make a prediction). Next for Neural networks there is a neural tree constructed which sets itself for each input and changes each time. It doesn’t perform well when we have large data set because the required training time is higher. It also doesn’t perform very well, when the data set has more noise i.e. target classes are overlapping. SVM doesn’t directly provide probability estimates, these are calculated using an expensive five-fold cross-validation. It is included in the related SVC method of the Python scikit-learn library.

3.2 Data Collection and Performance Metrics

3.2.1 Data Collection

The datasets used in the experiment are collected from MIB projects. It consists of Genuine and Fake Twitter datasets. The Genuine accounts dataset contains accounts of people who came forward to be part of academic study for detecting fake accounts on Twitter and it is mostly a mixture of accounts of researchers, social experts and journalists from Italy, US and other European countries. The fake accounts were purchased from three different Twitter online markets namely fastfollowerz.com, intertwitter.com and twittertechnology.com

3.2.2 Evaluation Metrics

In order to evaluate the performance of the system, various evaluation metrics are used based on following four standard indicators

- True Positive (TP): True positives are records that are correctly detected with expected vectors.
- True Negative (TN): True negatives are records correctly detected expected as Neutral.
- False Positive (FP): False positives are records that were detected by the system as expected but are listed in the other vectors.
- False Negative (FN): False negatives are records not detected by the system.

The evaluation metrics considered are

1. Accuracy which gives the ratio of number of correct results to the total number of inputs
2. Precision which gives the proportion of positive detection that was actually correct
3. Recall which gives the proportion of actual positives that was detected correctly
4. F1 Score which takes into account both precision and recall to compute the score. F1-score is given by harmonic mean of precision and recall. If F1-score is 1, then it is best value and worst is 0.

Machine Learning Algorithm	TN %	TP %	FN %	FP %	Precision %	Recall %	F-Measure %
Random forest	94.69	94.20	17.45	3.76	96.16	71.04	81.71
Decision Tree	82.57	88.90	33.64	5.71	93.96	67.04	78.25
Naïve Bayes	79.24	81.01	57.95	7.19	91.85	61.09	73.38
Neural Network	78.17	89.33	32.57	7.54	92.21	67.36	77.85
SVM	55.64	96.58	10.45	14.92	86.62	72.83	79.13

Fig 3.2.2.1 Evaluation Metrics

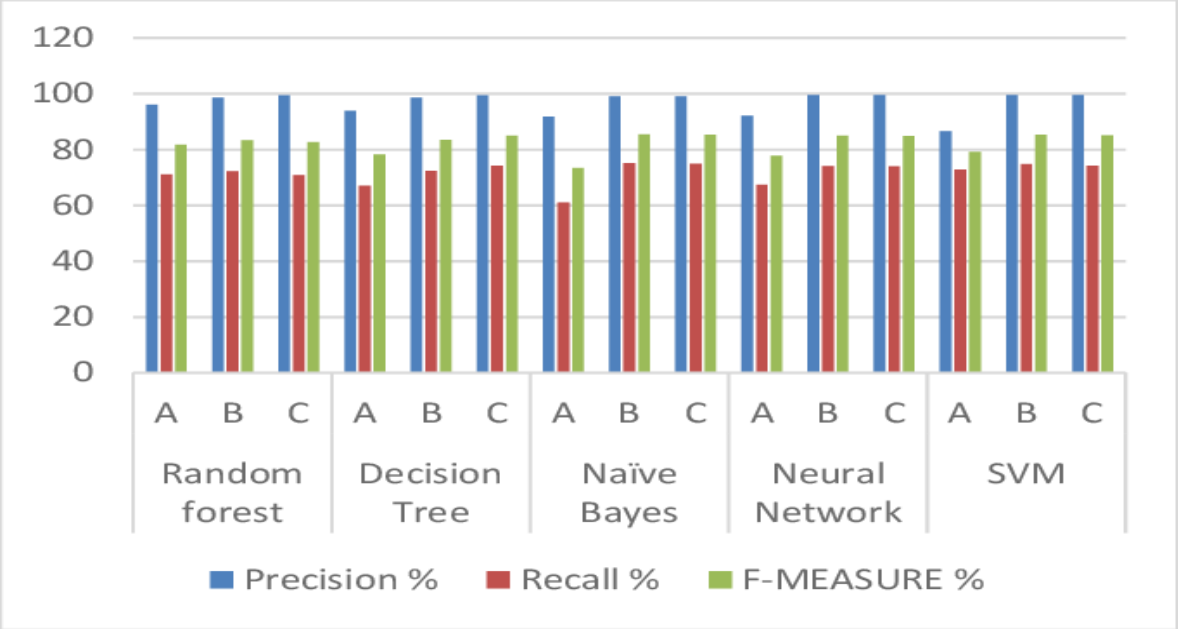

Fig 3.2.2.2 Graph for precision recall f-measure of different algorithms

CHAPTER 4
CONCLUSION

CHAPTER 4
CONCLUSION

Fake and clone profiles have become a very serious problem in online social networks. We hear some or the other threats caused by these profiles in everyday life. So, a detection method has been proposed which can find both fake and clone Twitter profiles. For fake detection, a set of rules were used which when applied can classify fake and genuine profiles.

CHAPTER 5
REFERENCES

CHAPTER 5
REFERENCES

- Sowmya P and Madhumita Chatterjee, ”Detection of Fake and Cloned Profiles in Online Social Networks”, Proceedings 2019: Conference on Technologies for Future Cities (CTFC).
- Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis and Evangelos P.Markatos, “Detecting Social Network Profile Cloning”, 2013.
- Piotr Brodka, Mateusz Sobas and Henric Johnson, “Profile Cloning Detection in Social Networks”, 2014 European Network Intelligence Conference.
- Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angello Spognardi, Maurizio Tesconi, “Fame for sale: Efficient detection of fake Twitter followers”, 2015 Elsevier’s journal Decision Support Systems, Volume 80.
- M.A.Devmane and N.K.Rana, “Detection and Prevention of Profile Cloning in Online Social Networks”, 2014 IEEE International Conference on Recent Advances and Innovations in Engineering.
- Kiruthiga. S, Kola Sujatha. P and Kannan. A, “Detecting Cloning Attack in Social Networks Using Classification and Clustering Techniques” 2014 International Conference on Recent Trends in Information Technology.