ISSN No. 0976-5697

VOLUME 13 SPECIAL ISSUE 1, JUNE 2022

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Nobel Approach for Fake Profile Detection using different Machine Learning Algorithms

Anjali Vyas
Department of Computer Science and Engineering
Geetanjali Institute of Technical Studies
Udaipur, Rajasthan 313001, India
anjalivyas2502@gmail.com

Jaya Sisodiya
Department of Computer Science and Engineering
Geetanjali Institute of Technical Studies
Udaipur, Rajasthan 313001, India
jayapalasiya15@gmail.com

Bhavika Rajora
Department of Computer Science and Engineering
Geetanjali Institute of Technical Studies
Udaipur, Rajasthan 313001, India
bhavikarajora38@gmail.com

Payal Manghnani
Department of Computer Science and Engineering
Geetanjali Institute of Technical Studies
Udaipur, Rajasthan 313001, India
payalmanghnani11@gmail.com

Mohammad Adnan Sheikh
Department of Computer Science and Engineering
Geetanjali Institute of Technical Studies
Udaipur, Rajasthan 313001, India
adnansheikh2355@gmail.com

Abstract: Social activity of everybody in today's' generation has gotten associated with online social networks**.** These sites have made extreme changes in the way we follow our social lives. With the help of these sites making friends and stay connected has became easier. . As with the fast progress, there are also diverse effects have been taking place .People face the problems of fake profiles, false statics and so on. In this paper, we provide a framework for direct recognition of fake profiles .This framework uses various classification techniques in Machine Learning like SVM, Random Forest, K-nearest Neighbor, Decision tree Classifier to group the profiles into fake or genuine classes. It can be easily implemented online on social networks.

Keywords: Machine Learning, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbourand Social Media.

1. INTRODUCTION

Online social media is the place each person has a outlook then be able to keep connecting their relations, transfer their updates, join with the people having same likes [1]. Social networking sites are expanding rapidly and changing the way people stay in contact with each other. The online communities bring people with the same interests together which makes users easier to make new friends. The paper consists of two phases [1]. In the first phase, pre-processing of dataset and feature selection takes place. In the second phase, the pre- processed data gets trained by the Machine Learning algorithms.

2. PROPOSED METHODOLOGY

2.1Overview

Each profile in a social network consists of lots of information such as gender, no. of friends, no. of comments, education, work, etc. Some of this profile data is public and private. In our framework we have used only public information but when our framework will be used by the social networking companies itself then they can capture public as well as private information. **.** Social networking sites have their pros and cons. One of the major problems that social media users face is the presence of fake profiles which can harm their personal data. In this framework, the classification methods used are Support Vector Machine (SVM) [3], Random Forest, Decision Tree, K-Nearest Neighbor (KNN) to identify the real and fake profiles. These Classification Techniques are capable to detect and divide the fake and real profiles separately.[3]

The steps that we have followed for the identification of fake profiles are as follows.

1. First, all the features are selected on which the classification algorithm is applied.
2. After proper selection of attributes, the dataset of previously identified fake and real profiles are needed for the training purpose of the classification algorithm.
3. The attributes selected in step 1 are needed to be extracted from the profiles (fake and genuine).[4]

4. After this, the dataset of fake and real profiles are prepared. From this dataset, 80% of both profiles (real and fake) are used to prepare a training dataset and 20% of both profiles are used to prepare a testing dataset.
5. After the accumulation of the training and the testing dataset, the training dataset is feed to the classification algorithm.
6. The efficiency of the classifier is calculated by calculating the no. of correct predictions divided by total no. of predictions.

2.2 Proposed framework

Proposed system is embedded with various Machine Learning tasks and the architecture followed is as shown below. The proposed framework in shows chart the sequence of processes that need to be followed for continues detection of fake profiles with active learning from the feedback of the result given by the classification algorithm.

1. The detection process begins with the selection of the profile that needs to be tested.
2. After the selection of the profile, the suitable attributes (i.e. features) are selected on which the classification algorithm is implemented.
3. The attributes extracted is passed to the trained classifier. The classifier gets trained regularly as new training data is feed into the classifier.
4. The classifier determines whether the profile is fake or genuine.
5. The classifier may not be 100% accurate in classifying the profile so; the feedback of the result is given back to the classifier.
6. This process repeats and as the time proceeds, the no. of training data increases and the classifier becomes more and more accurate in predicting the fake profiles.

2.3 Classification

Classification is the process of learning a target function f that maps each record, X consisting of a set of attributes to one of the predefined class labels, Y. A classification technique is an approach of building classification models from an input data set. In our proposed system we have used multiple classification techniques.

Random-Forest Classification Technique:

Random Forest is one of the mostused machine learning algorithms, because its simplicity and the fact that it can be used for both classification and regression tasks.

It's asupervised learning algorithm [12]. As it's seen from its name, it's bagged decision tree models that split on a subset of features on each split; it creates a forest and makes it somehow random. The «forest» it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. For better understanding of the RF algorithm is necessary to explain what the main idea behind decision trees is. Depending on the features in each dataset, the decision tree model learns a series of questions to figure out the class labels of the instances. What makes this model successful is that it is non- parametric and it can handle heterogeneous data (ordered or categorical variables, or a mix of both). Furthermore decision trees fundamentally implement feature selection, making them at least to some extent robust to irrelevant or noisy variables and are robust to outliers or errors in labels [12].

To summarize, here is steps that Random Forest algorithm follows:

1 .Randomly chooses n samples from the training set with replacement.
2. Grow a decision tree from the n sample. At each node.
3. Repeat the steps 1 to 2 k-times.
4. Aggregate the prediction by each tree to assign the class label by majority vote.

Using Naïve Bayes Classifiers:

Naive Bayes classifiers belong to a family of simple probabilistic classifiers used in machine learning. These classifiers are based on applying Bayes theorem with strong (naive) independence assumptions between the features. Naive Bayes is a simple method for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

Table 1

Features	**Description**
Id	Id of user
name	User name
screen_name	Screen Name
statuses_count	Statuses Count
followers_count	Followers Count
friends_count	Friends count
favourites_count	Favourites Count
listed_count	Listed Count
created_at	Date of account creation_at
url	AcountUrl
lang	Language
time_zone	Time zone

location	Geographic Location
default_profile	Default profil
default_profile_image	Default Profil Image Status
geo_enabled	Géolocalisation
profile_image_url	Profile Image URL
profile_banner_url	Profile Banner URL
profile_use_background_image	Profile Background Image
profile_background_image_url_https	Profile background Image Url
profile_text_color	Profile Text Color
profile_image_url_https	Profile Image Url Https
profile_sidebar_border_color	Profile Sidebar Border Color
profile_background_tile	Profile Background Title
profile_sidebar_fill_color	Profile Side Bar Fill Color
profile_background_image_url	Profile Background Image URL
profile_background_color	Profile Background Color
profile_link_color	Profile Link Color
utc_offset	Offset Status
Protected	Protection Status
verified	Verification Status
Description	Description of Account
updated	Update Date

It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. Naive Bayes classifiers are a popular statistical technique of email filtering. They emerged in the middle of the 90s and were one of the first attempts to tackle spam filtering problem [11]. Naive Bayes typically use bag of words features to identify spam e-mail, an approach commonly used in text classification. Naïve Bayes classifiers work by correlating the use of tokens (typically words, or sometimes other constructions, syntactic or not), with spam and non-spam e-mails and then using Bayes theorem to calculate a probability that an email is or is not a spam message [11]. In our paper we will assess the impact of using Naïve Bayes classifiers in the prediction of fake or genuine profiles in social networks (Facebook Data set).

Classifiers :A decision tree is a popular classification method that generates tree structure where each node denotes a test on an attribute value and each branch represent an outcome of the test. The tree leaves represent the classes. This technique is fast unless the training data is very large. It does not make any assumptions about the probability distribution of the attributes value. The process of building the tree is called induction.

3. IMPLEMENTATION

3.1Dataset:
We needed a dataset of fake and genuine profiles. Various attributes included in the dataset are a number of friends, followers, status count.
Dataset is divided into training and testing data.
3.2Initial Features: After using the dataset, we go for feature selection phase. We observed that there was much of unneeded features, either have no meaning for our subject or full of NaN values, so to make our models train well, we decided to drop them and to let only those of will affect directly on the results.
3.3Features Selection: Feature selection is one of the basic concepts in machine learning which hugely impacts the performance of classification and prediction. In our work, and in order to make our models train well, we decided to use only features which will affect directly the results. The features on the final dataset were [13]: statues_countfollowers_count, friends_count and favorites_count.
Below is the meaning of each feature
Table 2.

Features	Description
statuses_count	Statuses Count
followers_count	Followers Count
friends_count	Friends Count
favourites_count	Favourites Count

4. EVALUATION PARAMETERS

Efficiency/Accuracy = Number of predictions / Total Number of Predictions Percent Error = (1Accuracy)*100

Confusion Matrix - Confusion Matrix is a technique for summarizing the performance of a classification algorithm. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making.

TPR- True Positive Rate

TPR=TP/(TP+FN) FPR- False

Positive Rate

FPR=FP/(FP+TN)

TNR- True Negative Rate

TNR=TN/(FP+TN) FNR- False

Negative Rate FNR=1-TPR

Recall- How many of the true positives were recalled (found), i.e. how many of the correct hits were also found.
Recall = TP / (TP+FN)

Precision- Precision is how many of the returned hits were true positive i.e. how many of the found were correct hits.
Precision = TP / (TP + FP)

F1 score- F1 score is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score.

ROC Curve- The Receiver Operating Characteristic is the plot of TPR versus FPR. ROC can be used to compare the performances of different classifiers.

5. RESULTS

The efficiency of the Random Forest Classifier in classifying data is 95%. We have taken 80% of the data for training dataset and 20% for the testing dataset.

6. CONCLUSION

In this paper, we provided an approach to identify the fake profile in social network . As we concluded in our paper, we demonstrate that with limited profile data our approach can identify the fake profile with 99.64% correctly classified instances and only 0.35% incorrectly classified instances, which is comparable to the results obtained by other existing approaches based on the larger data set and more profile Information. Our research can be a motivation to work on limited social network information and find solutions to make better decision through authentic data. Additionally, we can attempt similar approaches in other domains to find successful solutions to the problem where the least amount of information is available. In future work we expect to run our model using more sophisticated concepts such as ontology engineering, in order to semantically analyze user posts, and comportments. This later concept can improve the quality of prediction of fake or genuine profiles.

7. REFERENCES

[1]. Elyusufi, Z., Elyusufi, Y., Ait Kabir, M.: Customer profiling using CEP architecture in a Big Data context. In: SCA 2018 Proceedings of the 3rd International Conference on Smart City Applications Article No. 64, Tetouan, Morocco, 10–11 October 2018. ISBN: 978-1-4503- 6562-8

[2]. Patel, M., & Sheikh, R. (2019). Handwritten digit recognition using different dimensionality reduction techniques. International Journal of Recent Technology and Engineering, 8(2), 999-1002.

[3] .Ameena , A., Reeba, R.: Survey on different classification techniques for detection of fake profiles in social networks. Int. J. Sci. Technol. Manage. 04(01), (2015)

[4] H. Gupta and M. Patel, "Study of Extractive Text Summarizer Using The Elmo Embedding," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020, pp. 829-834, doi: 10.1109/I-SMAC49090.2020.9243610.

[5] H. Gupta and M. Patel, "Method Of Text Summarization Using Lsa And Sentence Based Topic Modelling With Bert," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 511-517, doi: 10.1109/ICAIS50930.2021.9395976

[6] Rao, K., Gutha, S., Raju, B. Detecting Fake Account On Social Media Using Machine Learning Algorithms. International Journal of Control and Automation. 13. 95-100 (2020).

[7] Sen, S., Patel, M., Sharma, A.K. (2021). Software Development Life Cycle Performance Analysis. In: Mathur, R., Gupta, C.P., Katewa, V., Jat, D.S., Yadav, N. (eds) Emerging Trends in Data Driven Computing and Communications. Studies in Autonomic, Data-driven and Industrial Computing. Springer, Singapore. https://doi.org/10.1007/978-981-16-3915-9_27.

[8] Bouckaert, R., Eibe, F., Hall, M., &Holnies, G., Pfahringer, B., Reutemann, P., Witten, I. (2010). WEKA— experiences with a Java Open-Source Project. Journal of Machine Leaming Research.

[9] Ameta, U., Patel, M., Sharma, A.K. (2021). Scrum Framework Based on Agile Methodology in Software Development and Management. In: Mathur, R., Gupta, C.P., Katewa, V., Jat, D.S., Yadav, N. (eds) Emerging Trends in Data Driven Computing and Communications. Studies in Autonomic, Data-driven and Industrial Computing. Springer, Singapore. https://doi.org/10.1007/978-981-16-3915-9_28

[10] Ramos-Pollàn, R., Guevara-López, M.A., Suérez-Ortega,C. et al. (2012) Discovering Mammography-based Machine LeamingClassifiers for Breast Cancer Diagnosis.

[11] Alsaieh, M., Alwif, A., Al-Salman, A., AlFayez, M., & Almuhaysin, A. (2014). TSD: Hetecting Sybil Accounts in Twitter. 2014 13th International Conference on Machine Learning and Naïve bayes,

[12] Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica (Ljubljana).

[13] Algorithm for Data Mining. International Journal of Advanced Research in ComputerScience and Software Engineering Volume 3, Issue 6, (2013 June).

