JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF FAKE INSTAGRAM PROFILES USING MACHINE LEARNING

I.Anupriya[1],V. Sowmiya[2], Dr. G. Devika,[3.]

Department of Computer Science[1,2,3], Mannar Thirumalai Naicker College[1,2,3],

Madurai-625004, TamilNadu.

ABSTRACT

The majority of people now use social networking sites as part of their everyday lives. Every day, a vast number of people build profiles on social networking sites and connect with others, regardless of their place or time. False identities play an important role in advanced persisted threats and are also involved in other malicious activities. Users of social networking sites not only profit from them, but they also face security concerns about their personal details. To assess who is promoting threats in social networks, we must first identify the user's social network profiles. It is necessary to differentiate between genuine and fake accounts on social media based on the classification.Detecting fake accounts on social media has historically focused on a number of classification methods. However, it is possible to boost the accuracy of fake profile identification in social media. Machine learning and technology is used in the proposed work to increase the percentage of fake profile prediction. In feature selection model chi-square algorithm is applied and choose best data. In classification method the various machine learning algorithms are implemented, the algorithms are Logistic Regression and Random Forest algorithm. The classification result based on accuracy, precision, recall, f1-score, sensitivity and specificity.

I. INTRODUCTION

In today`s Modern society, social media performs a essential function in everyone's life. The preferred motive of social media is to hold in contact with friends, sharing news, etc. The range of customers in social media is growing exponentially. Instagram has lately received significant reputation amongst social media customers. With greater than 1 Billion energetic customers, Instagram has turn out to be one of the maximum used social media sites. After the emergence of Instagram to the social media scenario, human beings with a very good range of fans were referred to as Social Media Influencers. These social media influencers have now turn out to be a go-to area for the commercial enterprise employer to market it their merchandise and services[1].The good sized use of social media has turn out to be each a boon and a bane for the society. Using Social media for on line fraud, spreading False records is growing at a speedy pace.[2][3]

Fake debts are the foremost supply of fake records on social media. Business corporations that make investments massive Sum of cash on social media influencers ought to recognize whether or not the subsequent received through that account is natural or not. So, there's a good sized want for a fake account detection tool, that could appropriately say whether or not the account is fake or no[4]t. In this paper, we use class algorithms in system gaining knowledge of to hit upon fake debts. The method of locating a fake account especially relies upon on elements including engagement charge and synthetic activity.

II. LITERATU RE REVIEW

Data from previous studies were combined to create this paper. Table 1 shows the comparative study of different methods.

TABLE I. COMPARITIVE STUDY

No.	Title	Year	Method
1	Prediction of Fake Instagram Profiles Using Machine Learning	2021	Using the combination of image detection and Natural Language Processing (NLP) to detect fake accounts on Instagram
2	Automatic Detection of Fake Profile Using Machine Learning on Instagram	2021	Using several Machine Learning algorithm to detect fake accounts such as ANN, Random Forest, and SVC
3	Classification of instagram fake users using supervised machine learning algorithms	2020	Using several Machine Learning algorithm to detect Insta-gram fake accounts such as Random Forest, Multilayer Perceptron, Logistic Regression, Naves Bayes, and J48 Decision Tree
4	Instagram Fake and Automated Account Detection	2019	Detection of accounts using several Machine Learning algorithms such as Naïve Bayes, Logistic Regression, SVM, and Neural Networks
5	Using Machine Learning to Detect Fake Identities: Bots vs Humans	2018	Using supervised machine learning models to train the datasets (SVM, rf, and Ada-boost)

III .METHODOLOGY

A. General Approach

- Collecting of data. The first issue that arises when attempting to resolve the issue of fake profile detection is the collection of user data[6].
- Figuring out the attributes that will be used by the classification algorithm.
- A sample of data that have already been classified to train the classifier. That is, we require a sample of profiles from fake and profiles from genuine. These data will be used to train the classifier[6].. The classifier's future performance will be more accurate the larger the sample.
- Using this sample, training the classifier. Fig 1 shows this process.

B.Chi-square algorithm

A chi-square (χ^2) statistic is a test that measures how a model compares to actual observed data. The data used in calculating a chi-square statistic must be random, raw, mutuall yexclusive, drawn from independent variables, and drawn from a large enough sample. For example, the results of tossing a fair coin meet these criteria.Chi-square tests are often used to test hypotheses[7]. The chi-square statistic compares the size of any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship.

$\chi c2=\sum Ei(Oi-Ei)2$

where :c=Degrees of freedom;

O=Observed value(s);

E=Expected value(s).

C.Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes[8]. In simple words, the depen\\dent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

D.Random Forest algorithm

"Random Forest is a classifier that contains a number of decision trees on various subsets of thegiven dataset and takes the average to improve the predictive accuracy of that dataset."[9] Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

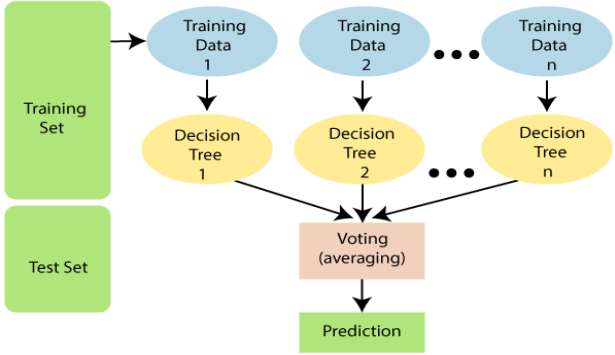

Fig. 1. General Process

E. Natural language processing

Natural language processing (NLP) is a field that focuses on making natural human language usable by computer programs. **NLTK**, or natural language toolkit, is a Python package that you can use for NLP.A lot of the data that you could be analysing is unstructured data and contains human-readable text[11]. Before you can analyse that data programmatically, you first need to pre-process it. In this tutorial, you'll take your first look at the kinds of **text pre-processing** tasks you can do with NLTK so that you'll be ready to apply them in future projects. You'll also see how to do some basic **text analysis** and create **visualizations**.

IV. CLASSIFICATION

A. Data Selection and loading

The process of loading the transformed data where users can access it is known as Data Load. Loading is a two-step process if the architecture has a staging database. The first step is to load transformed data into the staging database. Transfer the data to the warehouse or market from the staging database.

B. Data pre-processing

Preparing raw data for use in a machine learning model is the goal of data pre-processing. When creating a machine learning model, this is the first and most important step.When starting a machine learning project, clean and properly formatted data are not always available[12]. Additionally, data must be formatted and cleaned before being used in any way. Thus, we use the data pre-processing task for this

C. Model Selection

In machine learning, model selection is choosing the best model for our data. On different data sets, the performance of various models will vary, possibly significantly. Additionally, distinct models may offer distinct advantages. we can, for instance, use Logistic Regression to learn the coefficients of the model and explain how each feature affects the final prediction. Like Logistic Regression's coefficients, bagged tree models like Random Forest can tell you the Feature Importance of each column in the model

D.Classification

Based on the training data, the Classification algorithm is a Supervised Learning method used to classify new observations[13]. A program performing classification divides a new observation into a number of classes or groups after

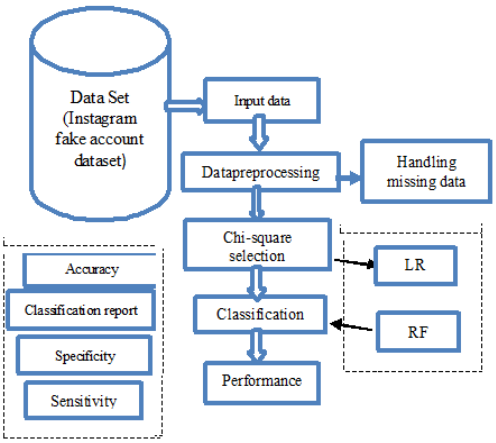

Fig. 2. Classification Process

learning from an existing dataset or observations. such as cat or dog, yes or no, 0 or 1,spam or not, etc. Targets, labels, or categories are all possible names for classes.

E. Performance Estimation

Different metrics, also known as performance metrics or evaluation metrics, are used to assess the model's quality or performance. We can gain a better understanding of how well our model has performed

TABLE II. COMPARITIVE ANALYSIS

Parameter	Chi-square algoritm	Linear Regression	Random Forest	Natural language processing
Testing data score	0.954	0.551	0.997	0.889
Accuracy	0.946	0.542	0.966	0.851
AUC	0.98	0.434	0.98	0.85
RMSE	0.251	0.331	0.286	0.381
MSE	0.065	0.119	0.083	0.148

with the given data by using these performance metrics. By tuning the hyper-parameters in this manner, we can enhance the performance of the model. Performance metrics assist in determining how well a machine learning (ML) model can generalize to new or unseen dataset. A comparative analysis of result obtained by using various algorithms is shown in the table above

F. Prediction

Prediction is the process of finding the numerical data for a new observation that is either missing or unavailable. In prediction, a predictor's accuracy is determined by how accurately it can predict the value of a predicated attribute from new data. An ordered value or continuous-valued function-predicting model or predictor will be developed. Predicting the best course of treatment for a particular disease, for instance, is an example of prediction.

V. CONCLUSION

We combined natural language processing methods with machine learning algorithms in this paper. We can easily identify fake profiles on social networking sites using these methods. We used the dataset in this paper to find the fake profiles. To find the fake profiles, the dataset and the machine learning algorithm dataset are analysed using NLP pre-processing techniques. The dataset and machine learning algorithm are analysed with the help of the NLP pre-processing methods.

VI.FUTURE WORK

A fully automated public Turing test, Captcha is a computer test used to distinguish between a human and a computer system user.SMS verification involves sending a user a confirmation code via SMS to his phone number, which he is required to enter in the appropriate field when registering or logging in to the system. A rate limit (also known as a bandwidth limit) restricts the number of system requests for a predetermined period of time.

REFERENCES

[1] A. M. Vegni, V. Loscri, A. Benslimane, SOLVER: A Framework for the Integration of Online Social Networks with Vehicular Social Networks, in: IEEE Network 34(1), 2020,pp.204-213,doi: 10.1109/MNET.001.1900259.

[2] Ml-cheatsheet.readthedocs.io. (2019). Logistic Regression — ML Cheatsheet documentation. [Online] Available at: https://mlcheatsheet.readthedocs.io/en/latest/logisticregression.html#binarylogistic-regression [Accessed 10 Jun. 2019]

[3] A. U. Hassan, et al., Sentiment analysis of social networking sites (SNS) data using machinelearning approach for the measurement of depression, in: 2017 International Conference on Information and Communication Technology Convergence (ICTC), Jeju,2017,pp.138140,doi:10.1109/ICTC.2017.8190959

[4] D. Ageyev, et al., Infocommunication Networks Design with Self-Similar Traffic, in: IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), Polyana, Ukraine, 2019,pp.24-27,doi: 10.1109/CADSM.2019.8779314.

[5] Z. Hu, et al., Development and Operation Analysis of Spectrum Monitoring Subsystem 2.4–2.5GHz Range, Lecture Notes on Data Engineering and Communications Technologies, 2020, pp. 675–709. doi: 10.1007/978-3-030-43070-2_29.

[6] Akyon, F. C., & Esat Kalfaoglu, M. (2019). Instagram Fake and Automated Account Detection. Proceedings- 2019 Innovations in Intelligent Systems and Applications Conference, ASYU 2019. https://doi.org/10.1109/ASYU48272.2019.8946437

[7] Dey, A., Reddy, H., Dey, M., & Sinha, N. (2019). Detection of Fake Accounts in Instagram Using Machine Learning. International Journal of Computer Science and Information Technology, 11(5), 83–90. https://doi.org/10.5121/ijcsit.2019.11507

[8] Meshram, E. P., Bhambulkar, R., Pokale, P., Kharbikar,K., & Awachat, A. (2021). Automatic Detection of Fake Profile Using Machine Learning on Instagram. International Journal of Scientific Research in Science and Technology, 117–127.https://doi.org/10.32628/ijsrst218330

[9] Sheikhi, S., 2020. An Efficient Method for Detection of Fake Accounts on the Instagram Platform. Revued Intelligence Artificielle, 34(4), pp.429-436.

[10]"Hidden Layer Definition" https://deepai.org/ma-chine-learning-glossary-and-terms/hidden-lay-er-machine-learning

[11]W. Delu, Enterprise Network Marketing Strategy Based on SNS Social Network, in: 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA), Xiangtan, China,2019,pp.295299,doi:10.1109/ICICTA49267.2019.00069.

[12]Y. Romanyshyn, et al., Social-communication web technologies in the higher education as means of knowledge transfer, in: IEEE 2019 14th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), 2019, pp. 35–39.

[13]M. Zharikova, V. Sherstjuk, Academic integrity support system for educational institution, in: 2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON, 2017.