Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 4; April-June, 2016, pp. 319-321
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Detection of User Identity in Social Networks

Satinder[1], Sanjeev Dhawan[2] and Kulvinder Singh[3]

[1]*M.Tech. (Computer Engineering), Department of Computer Science & Engineering, University Institute of Engineering and Technology (U.I.E.T), Kurukshetra University, Kurukshetra, Haryana*
[2]*Faculty of Computer Science and Engineering, Department of Computer Science & Engineering, University Institute of Engineering and Technology (U.I.E.T), Kurukshetra University, Kurukshetra, Haryana*
[3]*Faculty of Computer Science and Engineering, Department of Computer Science & Engineering, University Institute of Engineering and Technology (U.I.E.T), Kurukshetra University, Kurukshetra, Haryana*
E-mail: [1]satidersaini889@gmail.com, [2]rsdhawan@rediffmail.com, [3]kshanda@rediffmail.com

Abstract—*In today's world, social media is used every individual for expressing their feelings, opinion, experiences and emotions. Applying data mining on all these emotions expressed in posts, comments and likes called as social media data. This study aims to verify the owners of social accounts, in order to eliminate the effect of any fake accounts on the people. This paper targets to detect the genuine accounts versus fake accounts by using writeprint identification method which is writing style biometric. There are various strategies which are used by the researchers to detect the authorship identification. Here, in this research paper Supervised Machine learning technique is used. The data which has been provided for the entire process to be performed is extracted from the Twitter with the help of R package as it provides interface with the Twitter web API (Application Programmable Interface). Various Calculations has been performed to calculate the accuracy, precision, and other parameters.*

Keywords*: Social networks, identity recognition, text mining, R package*

1. INTRODUCTION

Recognizing the identity of the user in social networks is arising issue because of the substantial increase in the number of fake accounts, parody accounts and the accounts used for hate speech and violence. This problem causes moral and legal issues that affect the social life of innocent people. Celebrities of media and politics domains, for example, are regular victims of electronic identity framing. The increasing rate of using social networks contributed in boosting this problem up due to the simplicity of creating fake accounts. Many researches attempt to introduce new methods for identity recognition to overcome this problem. Research in 'identity recognition' intersects with many fields of science including biometrics, text mining, pattern recognition and social networks. This paper compares the performance of several classifier algorithms on a standard database for accuracy, precision, and other parameters. The data which has been provided as input to this process is extracted from Twitter with the help of R package. R package is a tool which is very much used for the statistical computing and graphics [1]. On the data which is extracted by R package further steps are performed including Pre-processing, Feature Extraction and then evaluation. At the end of this paper the results along with the conclusion and future scope are given and a system which is time efficient and accurate is introduced.

2. LITERATURE REVIEW

The present literature survey concentrates on the work done by a number of researchers worldwide in the field of recognize user identity in social networking websites. Fingerprint based identification has been the oldest biometric technique successfully used in conventional crime investigation. The unique, immutable patterns of a fingerprint- the pattern of ridges and furrows as well as the minutiae points-can help a crime investigator infer the identities of suspects. The absence of fingerprints in cyberspace leads law enforcement and intelligence community to seek new approaches to trace criminal identity in cyber-crime investigation. To overcome this problem they proposed new approach called 'writeprint' hidden in people's writings. Similar to a fingerprint, a writeprint was composed of multiple features, such as vocabulary richness, length of sentence, use of function words, layout of paragraphs, and keywords [2]. This approach was developed by R.Zheng *et al*. (2006) [3] to help identify an author in cyberspace. They also developed a GA-based feature selection model to identify the key features of writeprint for online messages and the identified key features achieve more comparable, higher accuracy and effective differentiate the writeprint of different online authors. Furthermore, they also applied the writeprint identification approach to other problems like intellectual property checking and plagiarism detection. Writeprint is the technique of predicting the most likely authorship of anonymous text by using stylistic information hidden in texts .They also developed a framework for authorship identification of online messages to address the identity tracing problems. They conduct experiments on English and Chinese online news group messages. There were some approaches to reduce the problem of author identification. The one was writer-independent approach presented by Pavelec D *et al.* [4] which reduces the problem of author identification to one model with two classes, which makes it possible to build a robust identification system using few genuine samples per author. The one major approach was Support Vector Machine for author identification was introduced by Joachim D *et al.* [5]. It was especially suited for this task as the feature spaces has very high dimensions, most features carry important information and the data for specific instances was sparse. SVMs for authorship attribution and text mining can process documents of significant length and databases with a large number of texts. SVM technology was firmly grounded in computational learning theory and training times compare favourably with other methods such as neural networks. Therefore, SVMs are currently the method of choice for authorship attribution. Stamatatos E [6] described the solution for the problem of author identification. The Common N-Gram (CNG) method was a language-independent profile-based approach with good results in many author identification experiments so far. This introduced approach provides a more stable solution than traditional CNG for high values of profile length. This was particularly important, especially in cases where there are only limited training texts for at least one of the candidate authors. Z. Liu *et al.* 2013 [7] proposed another approach for writeprint identification was semi-random subspace. Unlike conventional random subspace method which completely randomly select features from the whole feature space, semi-random subspace takes the account the distribution of individual-author writeprint hidden in feature space as well as make full use of the discrepancy among individual writeprint. They first divided the whole feature set into several individual author feature set (IAFS) in a deterministic way, then constructed a set of base classifiers on different randomly sampled feature sets from each IAFS, and finally combined all base classifiers for the final decisions. Researchers believe that the writing style of each person has unique linguistic features. Being able to extract these features will highly increase the chance of identifying fake accounts in social networks. This will help to overcome malicious activities against social network users. This was the research work which have been studied in the context of the user identification and on the basis of the all the information it is concluded that the number of problems have been solved with all the techniques and number of algorithms are available for the user identification.

3. IMPLEMENTATION

The work is performed in five steps namely; extracting the data, Pre-processing, Feature Extraction, Classifier and Evaluation. Extraction involves the use of R package named tool which helps to get the data from the Social Networking Website Twitter, Pre-processing is the method which is to be performed on the data provided so that the complexity can be reduced. After this Feature Extraction is performed and for this certain parameters has been set and according to those parameters whole data gets checked. Later, classifiers are applied and then evaluation is performed.

3.1 Extracting the data

R package is the tool which is used to extract the data from Twitter. This tool is basically used for statistical Computing and graphical display. It is free environment software. It runs on windows, Linux and Mac OS. R can be easily extended with 6,600+ packages available on CRAN.R package in this research work has been used with the purpose of extracting the tweets from the Twitter Website. R package is an interface which is used with the Twitter Web API.

3.2 Pre-processing

This is performed to reduce the complexity of the data provided by making it simple and easy to read.

3.3 Feature Extraction

This is the process in which number of features set is maintained on the basis of which input is examined. In this work , the set maintained is combination of four features and that is, total number of words in a line (word count), no. of URL (url) , no. of URL present per word count (URL/word count), Retweets (rt).

For example: Input: RT @BBCSport: Stuart McCall appointed #Rangers manager until the end of the season http://t.co/oJRIISNwYI.

3.4 Classifiers

KNN classification is used which is basically a non-parametric method and used for classification and regression. It depends on the output of the system whether KNN is used for the classification or for regression.

The K-Nearest Neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small).

Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis.

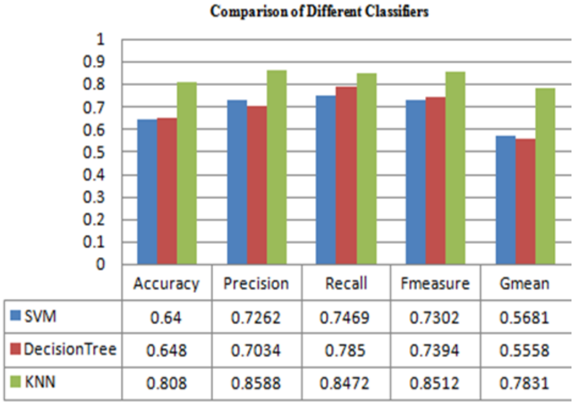
The process of constructing lexical ontology by analysing unstructured text is termed as ontology refinement by decision tree. Different algorithms of decision tree are used for classification in many application areas, like financial analysis, astronomy, molecular biology, and text mining.

3.5 Evaluation

On the basis of the process performed above, Accuracy, Time Complexity, F-measure, Recall, G-mean, Precision are evaluated and then the corresponding graphs get generated.

4. RESULTS

The data which is provided for entire process is extracted from social networking website "Twitter" using machine learning techniques. The Fig. 1-2 shows the comparison between different classifiers to optimize the Accuracy, Precision and Time Complexity and other factors.



	Accuracy	Precision	Recall	Fmeasure	Gmean
🟦 SVM	0.64	0.7262	0.7469	0.7302	0.5681
🟥 DecisionTree	0.648	0.7034	0.785	0.7394	0.5558
🟩 KNN	0.808	0.8588	0.8472	0.8512	0.7831

Fig. 1 Comparison of different Classifiers

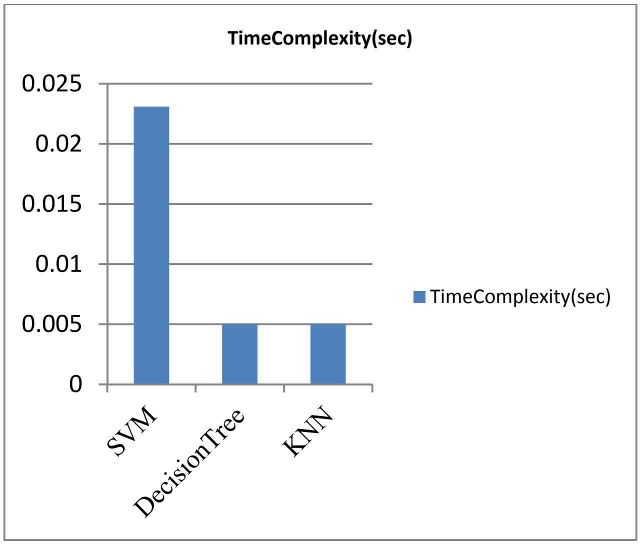


Fig. 2 Time Complexity

Entire work is performed in MATLAB which is abbreviated as Matrix Laboratory, it is used to perform the data visualization, data analysis and it is a very high-level interactive programming language [8]. Different classifiers are implemented in MATLAB for analyse the Accuracy, Precision, Recall, F-measure, G-mean, and Time Complexity.

For performing this work, a dataset of 150 records has been maintained and those records are basically the tweets, which are downloaded from Twitter using the R package tool and that data is saved as tdata.csv, here extension .csv means "comma separated values" A Comma Separated Values files stores tabular data in plain text.

5. CONCLUSION AND FUTURE SCOPE

Different classification techniques are used to optimize the accuracy, precision, time complexity, and other factors for a given input. On the basis of graphs obtained, it can be concluded that the KNN classifier is giving best results as compared to other two classifiers. In future there is need to increase the testing dataset and researchers can do the work of user identification by using the different classifiers or can introduce an entire new methodology by using any other classification techniques.

REFERENCES

[1] Data Mining with R, http://www.rdatamining.com/docs/introduction-to-data-mining-with-r Accessed on 14-may-2016

[2]J. Li, R. Zheng, and H. Chen,'' From fingerprint to writeprint,'' Communications of the ACM (2006); 49(4): pp.76-82

[3] Zheng. R, Li. J, Chen H, Huang Z,'' A framework for authorship identification of online messages: writing-style features and classification techniques,'' Journal of the American Society for Information Science and Technology, February 2006; 57(3): pp.378-393.

[4] Pavalec D., Justino E., and Oliveira L.S.," author identification using stylometric features," Artificial Intelligence, 2007; 11(36):pp.59-65

[5] J. Diederich, J. Kindermann, E. Leopold, and G. Paass," Authorship attribution with Support Vector Machines," In: Journal Applied Intelligence, 2000; 19(2):pp. 109-123.

[6] E. Stamatatos," Author identification using imbalanced and limited training texts," In: Proceedings of the 4th international workshop on text-based information retrieval, 2007:pp. 237-241.

[7] Z. Liu, Z. Yang, S. Liu, and Y. Shi," Semi-random subspace method for writeprint identification," Neurocomputing, May 2013; 108:pp.93-102.

[8] MATLAB, http://in.mathworks.com/products/matlab/ Accessed on 20-May-2016.