Architecture for User's Identification on Social Media Using Writeprint based on Distributed Machine LearningTechniques

Hardik Ashar
Independent Researcher
Mumbai, India.
dr.hardikashar@gmail.com

Abhishek Murkute
Independent Researcher
Mumbai, India.
abhishekmurkute@yahoo.com

***Abstract*—From dawn of social networking sites, there has been an instantaneous growth in its popularity. It is because of this reason several issues has emerged. This has allure the cyber criminals. An illicit use of publishing online messages or blogs for illicit intent has become an issue of great concern. This has abundantly assist to a ever precarious threat to social networks. Therefore, user identification is necessary.**

Biometric information is of two types scilicet biological and behavior. The writeprint falls in behavioral type of biometric information. This information is used for further analysis and detection purpose. This paper is regarding improvement on previously proposed framework designed for detecting user identification for social networking websites using a writeprint approach. In writeprint approach the semantic, stylometric, sentimental properties from written text of the user are captured, which is further used to detect user identity. Initially, the content written on social networking websites is downloaded. Later from the downloaded text, numerous features are extracted using text mining or web mining techniques. Most of these features are extracted natural language processing tool. Machine Learning techniques like Multi layered perceptron mode, Support Vector Machine (SVM), ensemble modeling methods in the distributed environment which is a supervised machine learning technique is used for user identity detection.

Keywords*— *writeprint; stylometric; multi layered perceptron model; user identification; social networking website, big data, machine learning;

I. Introduction

The cyber criminals have tendency to target the websites which have tremendous user visiting them. So that impact of the havoc created by them will be catastrophic. Social networking website can be defined as the web portal where group of individuals or organization meet, share and exchange their perspective and idea regardless of their different demography, profession, behavior, ethnicity etc. The social networking web sites have juggernaut users. Hence they are constantly under peril of falling victim to cyber terrorism. The social networking sites can be broadly divided into two group one, general purpose social networking websites like Twitter, Facebook. Second is specific purpose social networking websites like Academia.edu, aNobii or Virb.

In specific purpose social networking website is visited only by people or organization having interest certain or limited number of things. For example academia.edu is the website where people who are interested in research areas visit. The Virb, is the social networking website specifically for artist including photographer, musicians etc to share their ideas.

Anobii is social networking website especially for bibliophiles. Antithetically to specific purpose social networking website, general purpose social networking sites have all type of organization or group of people visit and are very popular. But social networking sites are more prone to cyber attacks than specific purpose social networking sites, as they have tremendous users visiting them all the time. Twitter, Facebook, Orkut and Google plus are some of the well known social networking websites. There are various issues of social networking websites such as cyberstalking, security, social profiling and third party disclosure.

The most vital of all issues for social networking website is security issue. The cyber convicts exploit the social networks to sending or publishing iniquitous contents, messages or text for illegal purposes. It has tremendous ruinous effect on society. Hence, it is necessary to curb such malicious activities. The identity recognition of the user or author provides a great aid in such cases. These kinds of threats to social networking sites have given rise to a new concept known as writeprint to keep check on the various malicious activities regarding authorship or user identification of the malefic text. The authorship identification falls in category of classification problem. Writeprint literally means 'written content (typed text) in printed format' i.e. extracting the unique features from typed or printed text.

Primitively in crime, the culprit was being clenched using the fingerprint or handwriting analysis as evidences. This was insubstantial in cyberspace. Here the writeprint concept comes into the picture. The term writeprint was first introduced by Li in the year 2006 [1]. Every person has a distinctive style of writing. In the writeprint approach such unique features are analyzed and authorship recognition is done.

This paper proposes a improved framework for identifying user on social networking websites using a writeprint approach than the previous proposed one [9] . The proposed approach uses the concept of distributed systems. Three machine-learning techniques multi layered perceptron mode, Support Vector Machines(SVM) and ensemble modeling, which uses the various features extracted from tweets are used for user identification.

The remaining part of paper is structured as follows: Section 2, states current research and literature review. The section 3 explains the writeprint concept. Section 4, proposes system which uses the writeprint approach followed three machine learnong techniques. Section 5, explains the algorithm. Section 6, discuss the evaluation of proposed system. The section 7, discuses the results obtained. Section8, provides a concluding remark.

II. Background and Literature Review

Now-a-days there are currently solutions to detect the malefic activity on social networks which are tracked using IP and MAC address. The inconsistent or abnormal patterns are used to detect malicious activity on social networks. To keep check on such activities, yet such system are not enough to track hackers or cyber criminals. Since, proxy-server and the anonymizer are used by them which make it difficult or sometimes impossible to track them.

It is not practically possible to maintain the servers of the social networks in every state or a country. After a crime has occurred getting in detailed information for digital forensics is difficult since servers are not located in that county. To get details such as time of access, IP address and other such details from foreign country, the investigation has to go through a red tape. This delays the investigation process. The proposed system in this paper can be applied in such cases to avoid delay.

The writeprint concept involves various areas of computer science like soft-computing, natural language processing, social networks, data mining, stylometry, cyber-security and machine learning, distributed and high performance computing [2][3][4].The writeprint approach is tested using decision making algorithms like ID3 [1].These algorithms are used for training and generating classifiers to decide the categories of new vectors. Machine learning algorithm like Bayesian Multi Normal Regression (BMR) is applied in writeprint approach. The BMR approach is found to be accurate to implement the writeprint. User was identified on base of geographical condition, their background, age and gender [5]. The writprint approach can be used for various different foreign languages. The Multiple Probabilistic Reasoning Model (PRM) was used for identification of user in Chinese language. It was observed that PRM worked efficiently for Chinese language[6]. The various algorithms can be merged in the writeprint concept. Multi clustering algorithms like Expectation Maximization (EM), k-mean and bisecting K-means used. But it was observed that performance is decreased with increase in number of user [7]. Three machine-learning techniques viz. Support Vector Machines(SVM) and ensemble modeling and multi layered perceptron model, which a type of supervised machine learning neural network model is being used. Similarly the same features will be support vector machine. This combined input will be fed to the ensemble modeling to get better accuracy. These machine learning techniques will be deployed using high performance techniques like cloud , grid and cluster computing[17]

Fig 1. shows the previously proposed architecture. It used only the multi layered perceptron model, supervised machine techniques. It was able to give 53.84% accuracy. The drawback of the previously proposed framework was it comparative slow and accuracy was just above 50% [9] [16]. The distributed computing systems and multiple machine learning techniques will help to overcome these shortfalls time and space complexity as well as it help to get better accuracy than previous proposed system.

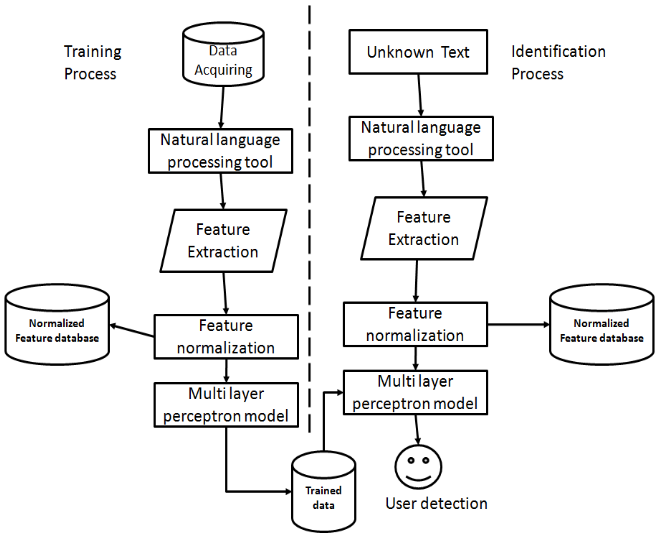
Training Process
Data Acquiring
Natural language processing tool
Feature Extraction
Normalized Feature database
Feature normalization
Multi layer perceptron model
Trained data
Unknown Text
Identification Process
Natural language processing tool
Feature Extraction
Feature normalization
Normalized Feature database
Multi layer perceptron model
User detection

Fig. 1.Previously Proposed Framework.

III. Writeprint

The writeprint concept states that every individual person has unique pattern of writing. The writeprint can be defined as a set of features that are extracted from the content written in printed form which are capable to recognizing the author of the content. Currently there are no standard protocol for the set features defined. Hence feature engineering plays a vital role. Every person in this universe has a unique style of writing depending on various factors such as personality, education, profession, command on vocabulary, nature, abbreviation etc. The person with aggressive personality would frequently use words like “must do”, “ought to”, “should”, etc. While person with calm personality or nature will use words like “may do”, etc.

The person with different profession will use some jargon for example person with commercial background will use words like depreciation, balance sheet. Person who is from teaching profession will use words like concepts, topic, marks. Person educated from computer science stream use worlds like algorithm etc. A statistician uses words like probability, percentage, and deviation, variance.

In research it was observed that a set of features differ from application to application. For example, we apply the proposed framework used for micro blogging website like twitter. The abbreviations like "wat" for "what"; ignoring vowels e.g. "hw" for "how"; number of hash and "@" tag while selecting features should be taken into consideration along with part of speech tagger, sentimental analysis which are extracted using natural language processing concepts. But the above mentioned feature might not be used other application. The features set can be divided in two broad type one primary features, second secondary features [8] [9]. The features like Number of hash tags used or Number of “@” symbol are features used in area of social net works. Example “how are? #abc #fgh @xyz”.

TABLE I. EXAMPLE

Sr. No	Tweets	No_word	No_pronoun	No_verb	No_hash
1	We won't forget these 2016 Moments any time soon. It was great. #BFF's	13	2	1	1
2	The best gift she ever gave me was her trust #wifey	11	1	2	1
3	I’m becoming convinced that the more hashtags a person uses, the less money that person makes…	16	1	3	0
4	Life, and what's happening in life,each and every little thing, be grateful it is in your life.	18	1	3	0
5	Do you think Ninja’s sneak up on their family members just for fun? #FunThoughts	14	1	2	1
6	She says "Three things are very important in life, to love, to live and to forgive"	16	1	3	0
7	To reach your greatest potential you'll have to fight your greatest fears #justdoit	13	1	2	1
8	One of my favorite things to do is to go for afternoon cycling. It feels great.	16	1	3	0

Consider an example in which only four features are used, Bob and Alice has a twitter account having user id say “2351”, “9873”. The four feature used are namely total number of word i.e. No_word, Number of pronouns i.e. No_pronoun, number of verb i.e. No_verb and total number of hash tags used i.e. No_hash. The multi layered perceptron model is trained for both Bob and Alice. There are eight tweets given in Table I, along with the four features extracted from it. tweets no: 1, 2, 5 and 7are of Alice, tweets no:3, 4, 6 and 8 are of Bob. First six tweets are used to trained by three machine-learning techniques viz. Support Vector Machines(SVM) and ensemble modeling and multi layered perceptron model and the last two viz. 7and 8 is used for testing purpose. In first six tweets observed that number of word Alice used in tweet written by Alice are in range 11-15. Number of pronouns used is in range of 1-2. Number of verbs used is range of 1-2. Number of hash tags he uses is 1. While the range of total words used Bob is 16-20, number of pronouns used is in range of 1-2. Number of verbs used are 3. Number of hash tags he uses is 0.

Suppose, the inference proposed is that, if more than two features the above inference the proposed framework will give user id of Alice i.e. “23510”. If it not matched then it will check for inference deduced by Alice. For Alice inference output will be user id of Bob i.e. “98730”. The user id is given as output because user id for every twitter account is unique and it also helps to handle case of people with same names. Hence for seventh tweet used id “23510” will be the output and for eighth it will give user id “98730”.

IV. Distributed System

The distributed system plays vital role in application using the writeprint approach. To speed up the process of high performance computing approaches like cloud computing, grid computing and cluster computing can be used as shown in the fig 2.

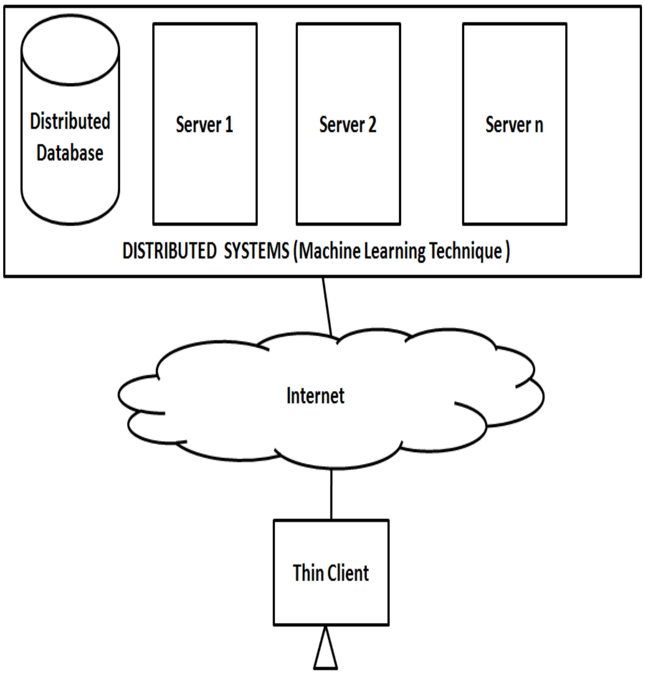

Fig. 2.Distributed Computing Architecture for Writprint Approach .

The distribute database will to help to store the tweets in the distributed database system. The servers will have various distributed machine learning techniques to help.

The recent technological development in distributed system and machine learning like Microsoft's azure machine learning, Google's tensorflow api for deep machine learning, even api like Apache spark. has provided boon

V. Proposed System

The proposed entire architectural framework will work on distributed computing environment like Hadoop, Apache spark, tensorflow. There are many social networking sites of which twitter is taken for the experimental purpose because it has 140 character limits for the tweets. The user can easily express thoughts in 140 characters, so maximum number of characteristics can be extracted. The proposed framework as shown in Fig. 3 consist of five phases: (1) acquiring tweets, (2) Feature extraction, (3) Normalization, (4) The distributed computing systems and multiple machine learning techniques, (5) User Identification. The detail of implementation for the proposed framework has been explained in detail which is shown in fig 1. The feature data base was given to the feature analysis using cosine similarity rule. These machine learning techniques will be deployed using high performance techniques like cloud , grid and cluster computing[17].

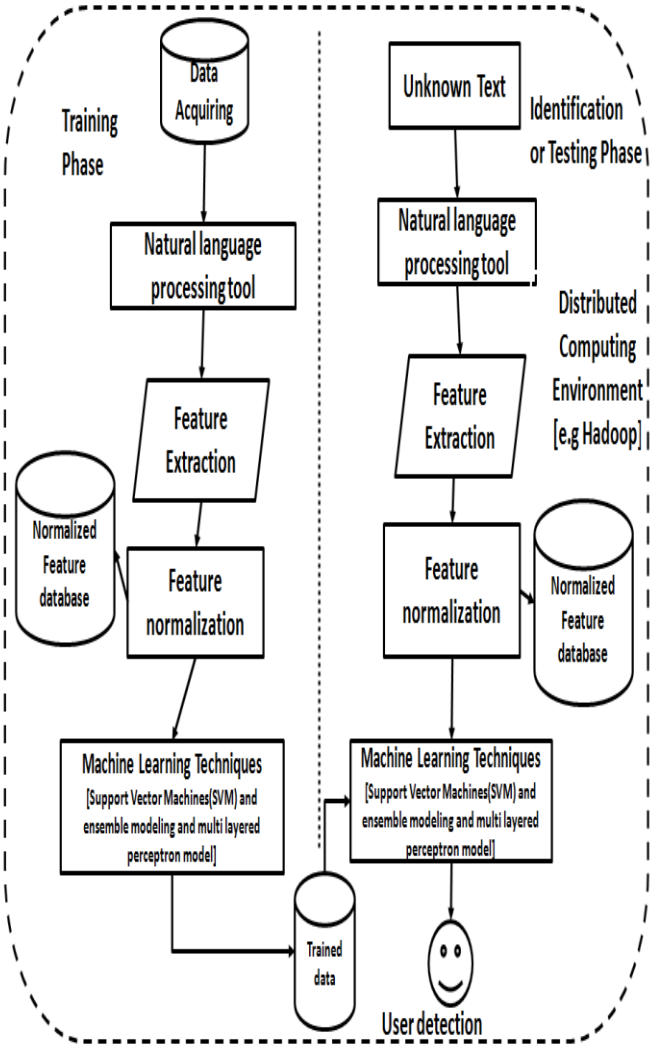
Training Phase
Data Acquiring
Natural language processing tool
Feature Extraction
Normalized Feature database
Feature normalization
Machine Learning Techniques [Support Vector Machines(SVM) and ensemble modeling and multi layered perceptron model]
Trained data
Unknown Text
Identification or Testing Phase
Natural language processing tool
Distributed Computing Environment [e.g Hadoop]
Feature Extraction
Feature normalization
Normalized Feature database
Machine Learning Techniques [Support Vector Machines(SVM) and ensemble modeling and multi layered perceptron model]
User detection

Fig. 3. New Proposed Framework.

A. Acquiring Tweets

The first stage is to acquire the tweets. There are various libraries available to download tweets from twitter. These api libraries are available to download tweets from twitter using various different technologies. Few of them are mentioned here. Asptwitter is used to download tweets using asp, TwitterJSClient in javascript, Twitter4j is java api library. The twitter4j java library is been used to download tweets [10]. The users were be selected from various different professions like sport, etc. The tweets of each user was been downloaded to in a comma-separated value format (*.csv). This data was used for further processing

B. Feature Extraction

The feature selection places vital role in writeprint approach. There are various natural language processing tools which can be used for part of speech tagger like Apache OpenNLP tool, Stanford Part-Of-Speech Tagger, GATE etc.

TABLE II. LIST OF FEATURES

Features	Description
No_noun	Number of nouns used
No_verb	Number of verbs used
No_participle	Number of participles used
No_pronoun	Number of pronoun used
No_preposition	Number of prepositions used
No_adverb	Number of adverbs used
No_conjunction	Number of conjunctions
No_words	Total number of words in the message
Ex_url	Percentage of external links used
No_hash_word	Percentage of # tags used
No_mentions	Percentage of @ tags used
Sentiments	The tweet has positive, negative or neutral sentiments
Smiley	The type of emotion smiley used
Is_ Smiley	Weather the user uses smiley
Era_Of_Words	The word used belong to which era i.e. ancient / medieval / modern English

The features was extracted from tweets by using libraries of Stanford Part-Of-Speech Tagger (POS). The Stanford POS Tagger is the tool which reads text in and assigns parts of speech to each word such as noun, verb. This tool is implemented in java [11]. Apart from total number of word used in tweets, number capital letters used in Table II, along with the description of features used during implementation.

C. Normalization

The features obtained are normalized. Normalizing the data means adjusting values measured on different scales to notionally common scale. The normalization is necessary to obtained better results. The normalization formula used is given in equation (1).

$$X' = a + \frac{(X - X\min)(b-a)}{(X\max - X\min)} \qquad (1)$$

Where X' is normalized output, Xmin= minimum value of data, Xmax=maximum value of data, a=minimum rage of normalized data, b= maximum range of normalized data.

D. Machine Learning Techniques

The three advance techniques will be deployed using high performance computing techniques like cloud, grid and cluster computing[17] [18] [19].

The machine learning techniques can be broadly classified as supervised and unsupervised machine learning techniques. The unsupervised machine learning techniques can be used as preventive measure. But for using writeprint approach as digital forensic tool the supervised machine learning techniques must be used as the accuracy of supervised machine learning techniques is better than unsupervised machine learning techniques. Here the supervised machine learning techniques are been used.

Multi layered perceptron model is being used for user identifying the user. In training stage, adequate amount of data should be use to train to get better results. The tanh activation function is used as its better output [12]. The figure 3 explains the role of multi layered perceptron model.

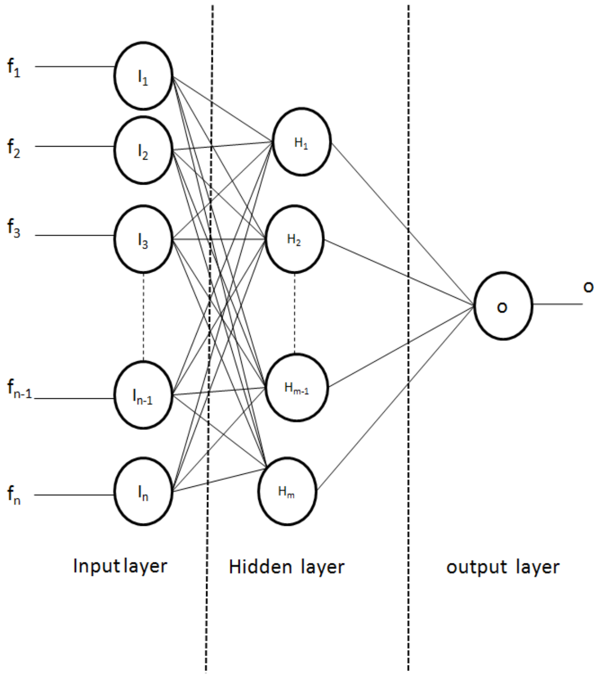

Fig. 3. Multi Layered Perceptron model

The f1, f2, …, fn are the input features to the neuron I1,I2, ..., In of input layer of multi layered perceptron model. The H1, H2... Hm are the neurons of the hidden layer. O is output neuron of output layer. Every edge has a weight assigned to initially it is near to zero. During the training process the weight will get updated. The activation function used is tanh. The output of the multi layered perceptron model is one of user's user id which used training phase [13] [14].

Along with the multi-layered perceptron model. the Support Vector Machine [SVM] will be use .The ensemble modeling will be use to increase the accuracy of the system.

The ensemble modeling is a most advance technique in machine learning where group two or more different machine learning models or techniques are used. Then synthesizing the results into a single score from which perdition or classification is been done.

E. User identification

This is the last phase here the output was generated the User Id and user name. User Id is unique to the every user account. Hence user id was been used. The cosine similarity was used performance evaluation parameter.

VI. Algorithm

The entire framework will be running on distributed computing environment like hadoop, apache spark, tensorflow etc.. Three machine-learning techniques viz. Support Vector Machines(SVM) and ensemble modeling and multi layered perceptron model, which a type of supervised machine learning neural network model is being used for the implementation and analysis of the framework is given below.

Algorithm:

1. Acquiring Tweet using twitter4j java library and storing them in csv format which is input to our system
2. Feature Extraction features are extracted using Sanford POS tagger, GATE as shown in table II .
3. Normalization of the features vector is done using equation (1).
4. Three machine-learning techniques viz. Support Vector Machines(SVM) and ensemble modeling and multi layered perceptron model.

   The Multi layered perceptron model technique is elaborated below:

   Input: N= Starting of neural network model. X={x_1….x_h} // Input tuple from training set. O= {O1} // Output tuple desired.

   Training: X is input layer. Initially weights small random values for edges Generate random weight values for edges connecting each input node i to hidden layer node j with weight $w_{i,j}$. Weighted sum of the input to the j^{th} node of the hidden layers is given by equation (2).

$$Net_j = \sum w_{ij} x_j \tag{2}$$

Based on the hidden layer generated, the actual output layer value which is calculated using activation function equation (3) [12] [13] [14].

$$x_k = \tanh(Net_j) \tag{3}$$

The difference between the actual output output x_k and the expected output O_k for which training is been done is given by equation (4).

$$\Delta_k = x_k - O_k \tag{4}$$

The error signal for node k^{th} in the output layer is calculated as equation (5).

$$\delta_k = \Delta_k * O_k(1 - O_k) \tag{5}$$

Modify the weight, $w_{i,k}$, between the output node, k, and the node j using learning rate, l_r (approx.=0.7) is given by equation (6) and equation (7).

$$w_{j,k} = w_{j,k} + \Delta w_{j,k} \tag{6}$$

$$\Delta w_{j,k} = I_r * \delta_k * x_k \tag{7}$$

Repeat the training step expect initialization till δ_k<0.05(approx) for every user.

Similarly the same features will be support vector machine. This combined input will be fed to the ensemble modeling to get better accuracy. These machine learning techniques will be deployed using high performance techniques like cloud , grid and cluster computing[17]

5. User Identification

   The steps 1 to 5 are executed based on Input tuples values generated from new posts captured by the algorithm where x_k is the input vector which gives the normalized feature value of user id (required for user identification on twitter) specified during training process.

This output was been compared by feature vector analysis and cosine similarity rule. Given as equation (8) [4].

$$\cos(P,Q) = \frac{\sum_{i=1}^{n} P_i x Q_i}{\sqrt{\sum_{i=1}^{n} (P_i)^2}\sqrt{\sum_{i=1}^{n} (Q_i)^2}} \tag{8}$$

Where P and Q are feature vectors.

VII. EVALUATION

The proposed system will be tested for the various numbers of unknown tweets. To test for accuracy of the proposed system, the equation (9) is used, as the performance evaluation parameter [15].

$$C.A = \frac{C.T}{T.T} \tag{9}$$

Where C.A=Classification Accuracy, C.T=Number of correctly identified tweets, T.T=Total tweets used for testing.

The accuracy proposed system based on the neural network model is compared with the feature vector analysis using the cosine similarity. The features extracted which are tabulated in the table II are used for computing the cosine similarity. These values are used to identify the user.

VIII. APPLICATIONS

The writeprint approach for can be used as both preventive approach and detection. It is said writeprint will be the future digital forensic tool of the digital epoch.

It will help the identify the anonymous author's of old scripts of various ancient literatures. It can help to check the online plagiarism.

This techniques will boon in area of medical science. It can help psychologist to understand the psychology of their patients.

IX. CONCLUSION

There are many serious issues of social networking websites namely cyberstalking, authorship identification, etc. Cyber criminals misuse the social networking website by writing iniquitous messages which intent to harm others. The proposed framework can be used to detect such crimes faster. This will reduce the time required for digital forensic because it help to bypass the red tape involved in investigation procedure.

The theoretical analysis shows that the proposed framework will improve user identification process. The factors that are taken in consideration are easy data acquisition, feature extraction. Here, in writeprint the feature engineering plays a vital role. Further the features extracted are normalized. The all three machine learning model should be trained with adequate amount of data.

The newly proposed system for detecting user identity on social networking web site using write print approach gives a better classification accuracy and reduce the time required for classification. the time to detecting use authorization will be reduce as distributed computing system like hadoop, Apache Spark, Azure Machine Learning, tensorflow is used.

X. FUTURE SCOPE

For the future work, more number of features can be added. The proposed approach has considered English language only. This can be future extended to other foreign languages and combination of two or more languages in a give statement e.g. " ‘शीलवृतफला विद्या ’ is the motto of Mumbai university!.". In future the approach can support detection the transliteration e. g." 'shilvrutaphala vidya', is the motto of Mumbai university ".

REFERENCES

[1] J. Li, R. Zheng, and H. Chen,"From fingerprint to writeprint.", Communications of the ACM - Supporting exploratory search. 76-82. Available: http://dl.acm.org/citation.cfm?id=1121951.

[2] Sara Keretna, Ahmad Hossny, Doug Creighton"Recognising User Identity in Twitter Social Networks via Text Mining", In the journal of Institute of Electrical and Electronics Engineers, 2013.

[3] Sara EL MANAR EL BOUANANI, Ismail KASSOU," Using lexicometry and vocabulary analysis techniques to detect a signature for web profile", In the journal of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013.

[4] Fredrik Johansson, Lisa Kaati, AmendraShrestha," Detecting Multiple Aliases in Social Media", In the journal of Institute of Electrical and Electronics Engineers, 2013.

[5] Shlomo Argamon, Moshe Koppel, James W. Pennebaker and Jonation Schler, “Automatically Profiling the author of an Anonymous text”, In communication of ACM , Vol 52,No:3, 2009.

[6] Hong Zhu, Zhaoli Zhang, Zhi Liu,"Online Writeprint Identification Via Multi-PRM", In the journal of Institute of Electrical and Electronics Engineers,2012.

[7] Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung and Mourad Dabbabi, “Mining writeprints from anonymous e-mails for forensic investigation”, In journal of Science Direct,2010.

[8] Ahemed Abbsi and Hsinchun Chen,"Writepeint:A Stylometreic Approach to Identitfication and similarity detection in cyberspace ",In the journel of ACM transaction on Information system, Vol 26,No:2, Article 7, 2008.

[9] Abhishek M. Murkute, Jayant Gadge, “Framework for User Identification Using Writeprint Approach”, IEEE International Conference on Technologies for Sustainable Development (ICTSD-2015), Feb. 04 – 06, 2015, Mumbai, India.

[10] Yamamoto, Yusuke. "Twitter4J-A Java Library for the Twitter API." (2010). Available: http://twitter4j.org/en, last visited on 04-11-2014.

[11] K.Toutanova Stanford Log-linear Part-Of-Speech tagger Available: http://nlp.stanford.edu/software/tagger.shtml.

[12] Barrey L. Kalman and Stan C. Kwasny, “Why tanh: Chooosing a Sigmodial funtion”, In the journal of Institute of Electrical and Electronics Engineers, 1992.

[13] Jacek M. Zurada, “Introduction to Artificial Neural System”, West Publishing Company, 1992.

[14] Tom M. Mitchell,"Machine Learning",McGraw-Hill Publication,1997.

[15] Ahemed Abbsi and Hsinchun Chen,""Sentiment analysis in multi languages: Feature selection for opnion classification", In the journel of ACM transaction on Information system, Vol 26,No:3, Article 12, 2008.

[16] Abhishek M. Murkute, Jayant Gadge, “ Identification of Users on Social Networking Website Using Soft Computing Technique ”,(presented in) IEEE International Conference on Technologies for Internet of things (IOTA-2016), Jan. 22 – 24, 2015, Pune, India.

[17] Kiranjot Kaur , Anjandeep Kaur Rai." A Comparative Analysis: Grid, Cluster and Cloud Computing ", International Journal of Advanced

Research in Computer and Communication Engineering Vol. 3, Issue 3, March 2014

[18] Zaharia, Matei, et al. "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing." *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 2012.

[19] Abadi, Martın, et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." *arXiv preprint arXiv:1603.04467* (2016).