iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** IV **Month of publication:** April 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Detection of Authenticity in Social Networks

Rajeshkumar Mourya[1], Varun Kulkarni[2], Rohit Kadam[3], Mukesh Rajpurohit[4]
[1,2,3,4] *Department of Computer Engineering, University of Pune, India*

Abstract— Social networks are the most convenient and effective means of communication in past few years. Our study aims to verify the owners of social accounts, in order to eliminate the effect of any fake accounts on the people. This study aims to differentiate between genuine accounts versus fake accounts by using writeprint identification method which is writing style biometric. We will first extract all the features using text mining techniques. The most important part of this process is machine learning in which we will train our system according to our usage and then we derive our knowledge database. From this database we will derive different vectors in accordance with the features extracted.
Keywords— text mining, identity recognition, social networks

I. INTRODUCTION

Data mining refers to extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as 'Knowledge mining from data' or "Knowledge Discovery in database".
Data mining consists of five major elements:

A. Extract, transform, and load transaction data onto the data warehouse system.
B. Store and manage the data in a multidimensional database system.
C. Provide data access to business analysts and information technology professionals.
D. Analyze the data by application software.
E. Present the data in a useful format, such as a graph or table.

This study aims to recognize the identity of the user on twitter social networking sites. Use of social networking sites has increased in tremendous amount and with security on high alert it is mandatory to determine the fake accounts and to put some restrictions on the usage. So we use different methods for identity recognition. This type of searching the author identification is also kown as "Write-print identification". The term 'writeprint' was first introduced in 2006 by Li. This method of identification is very much useful in unstructured data like emails and the one we are working on is Twitter.

II. BACKGROUND AND LITERATURE REVIEW

Every person has a unique style of writing, which depends on certain factors such as culture, education and the environment in which that person lives. We intend to extract suitable features and apply classification technique with the help of which a person can be easily identified using his texts. The process of extracting linguistic features from anonymous text with the aim of identifying the author of that text is called 'writeprint identification' [2].
Research in the field of writeprint identification has been done scarcely because of its tag of latest technological development in the field of computer science. The use of social networks has made it an intriguing topic. The term 'writeprint' was first introduced in 2006 by Li [3]. It was formally known as 'author identification'. Researchers use writeprint identification for analyzing different types of unstructured text like emails [4-5], online product reviews [6-7], news columns [2, 8], text messages, and chatting messages [9, 10].
Short text messaging services such as whatsapp, line etc. prove problematic for writeprint identification, this is because there is a less probability for extraction of the desired features due to insufficiency of text. Author identification is an archetypal type of classification problems. Relevant features need to be extracted with the help of different feature extraction techniques which define its accuracy; these extracted features are used as input to the learning module which produces a target category as an output.
Text classification techniques can generally be divided into three types: rule-based, statistical-based, and machine learning techniques. Each of these techniques can be applied to the lexical level, morphological level, syntactic level, or semantic level. If the domain of the text is limited or constrained, it is easier to analyse the text as explained in Figure 1.

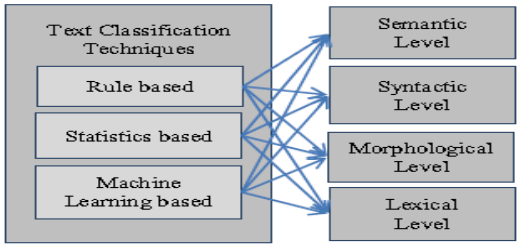


Fig. 1 Text classification techniques and their implementation on semantic, syntactic, morphological and lexical levels

III. CASE STUDY

A. *Case Study 1[]*

In this paper, a technique is proposed to detect writeprint identification by analyzing short anonymous text extracted from Twitter. In order to understand the unique identity of a user, the concept of machine learning technique is used. This is done by extracting certain features from a training data set of twitter accounts.

1) Data Collections: A Data crawler is needed to collect the information from different social networking sites. The crawler that used in our experiment is called Twitter4j. It is basically a software tool which is implemented with the help of a programming language Java. In this paper, we have considered 1000 twitter messages from 30 different accounts . The accounts have been selected particularly of celebrities, politicians and different famous figures. For a particular group of celebrities, 10 Accounts have been taken into consideration, so that analysis can be done on dissimilar varied results.

IV. PRE-PROCESSING

Nature of the tweets in the twitter social networking site is different than other social networking sites. For example, it does not support audio, video and embedding pictures in its tweets. But this type of media can be shared by attaching the link of that respective media to the text message. Yet another issue is that, it has limited size of the message, i.e. only 140 characters. Also people use different abbreviations such as “FYI” instead of For Your Information or “NYC” instead of New York City. Also many times people tend to remove certain vowels from a word such as – “what?” becomes “wt” and so on. One more issue is the absence or lack of punctuation and not to mention the use of special characters such as hash tags.

So there is need for the manipulation of these issues before the process of training and recognition.

V. FEATURE EXTRACTION

Feature extraction plays a very important role in the process of classification. Learning process and the final results are significantly affected by selection of appropriate set of features. Features that have been selected are listed in the following table I.

VI. LEARNING

Anonymous texts are classified using formulation of the classification model signifies the process of learning. Training of the learning model is done with the help of a set of twitter messages which are labelled with the user names. This learning process has been illustrated in figure 2, a feature vector is presented out of a set of features from different tweets.

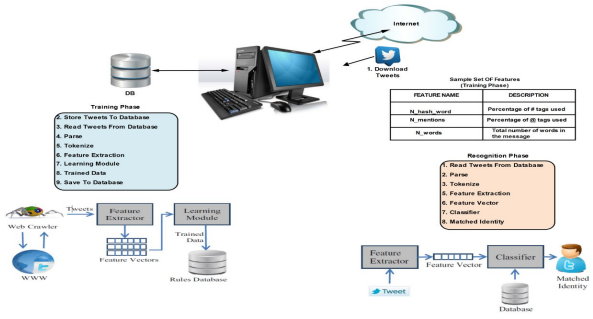


Fig.2 Proposed Architecture

Table 1: Set of features Obtained

Feature Name	Description
N_hash_word	Percentage of # tags used
N_mentions	Percentage of @ tags used
Ex_link	Percentage of external links used
Is_pic_exist	1, if images are linked to the twitter message, 0 otherwise
N_words	Total number of words in the message
Is_abbreviate	Does the message use abbreviations
Is_rem_vowels	Does the message contain summarized form of words
NE_types	What named entities are used in the message
N_noun	Number of nouns used
N_verb	Number of verbs used
N_participle	Number of participles used
N_interjection	Number of interjections used
N_pronoun	Number of pronouns used
N_preposition	Number of prepositions used
N_adverb	Number of adverbs used
N_conjunction	Number of conjunctions used
Freq_words	Top five most frequently used words
N_special_chars	Number of symbols like _ for example
N_capital	Number of capital letters

Using Twitter4j, a sample of tweets is extracted which signifies the learning. The training process is started by processing the samples of 50, 100, 150 and then 200 tweets. Comparison of the results is done to detect which sample size produces the best accuracy.

Example: A tweet from David Cameron by 13 May

TABLE 2
FEATURES EXTRACTED FROM TWEETS FOR THE TRAINING PHASE

Feature	Value	Description
N_hash_word	0	No # tags in this tweet
N_mentions	2	@whitehouse, @BarackObama
Ex_link	1	Pic.twitter.com/vCaoP9kwZI
Is_pic_exist	1	Pic.twitter.com/vCaoP9kwZI
N_words	16	@s, links and symbols are not counted
Is_abbreviate	1	US
Is_rem_vowels	0	Words are all spelled correctly
NE_types	[Country]	US
N_noun	3	US, phone-in, meeting
N_verb	4	Discuss, Keep, Updated, Plenty
N_participle	1	Doing
N_interjection	0	
N_pronoun	2	My, You
N_preposition	0	
N_adverb	1	Ahead
N_conjunction	2	Of, with
N_special_chars	2	In: 'phone-in' and 'discuss will'
N_capital	4	Capital letters in mentions and links are not calculated

VII. CLASSIFICATION

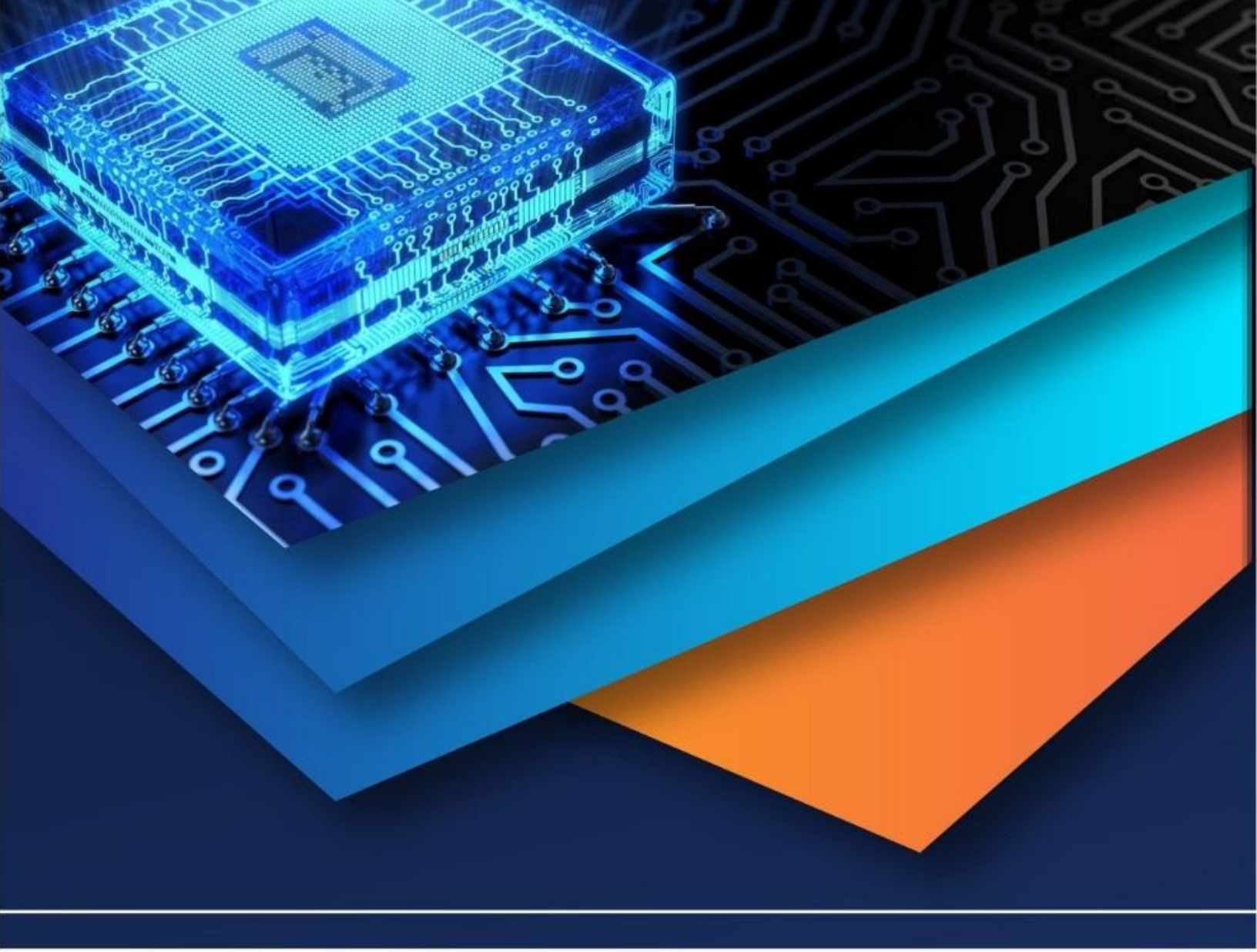
The process of classification starts when a message which is unlabelled is given as an input to the trained classification model. A particular set of features which is extracted from the message which is unlabelled forms a feature vector that goes to the classifier established in the learning stage. Then, the output of the classifier is the predicted author of the twitter messages. The process of classification has been illustrated in Fig. 2

The similarity between the feature vectors is measured using the Jacard's coefficient index. It is calculated in the following manner. Intersection of the two feature vectors is taken and it is divided by their union itself, as shown in equation 1.

$$J(v1, v2) = |v1 \cup v2| / |(1)$$

REFERENCES

[1] "Recognizing user identity in twitter social networks via text mining", Sara Keretna, Ahmad Hossny and Doug Creighton, 2013 IEEE International conference on systems, man and cybernetics.
[2] Z. Liu, Z. Yang, S. Liu, and Y. Shi, "Semi-random subspace method for writeprint identification, "Neurocomputing, vol. 108, pp. 93–102, May 2013.
[3] J. Li, R. Zheng, and H. Chen. (2006, April 2006) From fingerprint to writeprint. Communications of the ACM - Supporting exploratory search. 76-82. Available: http://dl.acm.org/citation.cfm?id=1121951
[4] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M.Debbabi, "Mining writeprints from anonymous emails for forensic investigation," Digital Investigation, vol. 7, pp. 56–64, Oct 2010.
[5] F. Iqbal, R. Hadjidj, B. C. M. Fung, and M. Debbabi," A novel approach of mining write-prints for Authorship attribution in e-mail forensics," in the Eighth Annual DFRWS Conference, Baltimore, MD, 2008, pp. S42–S51.
[6] J. Sun, Z. Yang, P. Wang, and S. Liu, "Variable Length Character N-Gram Approach for Online Writeprint Identification," in 2010 International Conference on Multimedia Information Networking and Security (MINES), Nanjing, Jiangsu, 2010, pp.486 - 490.
[7] S. Liu, Z. Liu, J. Sun, and L. Liu, "A Method of Online Writeprint Identification Based on Principal Component Analysis," in 2010 International Symposium on Information Science and Engineering (ISISE), Shanghai, 2010, pp. 319 - 321.
[8] E. F. Legara, C. Monterola, and C. Abundo,"Ranking of predictor variables based on effect size criterion provides an accurate means of automatically classifying opinion column articles,"Physica A: Statistical Mechanics and its Applications, vol. 390, pp. 110–119, Jan 2011 2011.
[9] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M.Debbabi, "A unified data mining solution for authorship analysis in anonymous textual communications," Information Sciences, vol. 231, pp. 98–112, May 2013.
[10] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, andF. Can, "Chat mining: Predicting user and message attributes in computer-mediated communication, "Information Processing & Management, vol. 44, pp.1448–1466, July 2008 2008.
[11] M. Verdone, "Python Twitter Tools (PTT)," 1.9.4 ed,2013.
[12] K. Toutanova. (2000). Stanford Log-linear Part-Of-Speech Tagger. Available: http://nlp.stanford.edu/software/tagger.shtml



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY