IDENTIFICATION AND PREVENTION OF MULTIPLE ACCOUNT IN SOCIAL MEDIA

E. Elangovan[1], Dr. D. Chandrakala[2]

[1]PG Student, [2]Professor

Computer Science and Engineering, Kumaraguru College of Technology,

Coimbatore, (India)

ABSTRACT

Identity deception has become an increasingly important issue in the social media environment .The past methods for detecting fake profiles have mainly focused on detecting deception through verbal communication (e.g., speech or text).Although these methods yield a high detection accuracy rate, they are computationally inefficient for the social media environment .The work concentrate on detection method based on non-verbal behaviour for identity deception ,which can be applied to many types of social media .The main goal is to yield a high detection accuracy rate and computational efficiency for the social media environment. The number of users registering with social networking sites such as Facebook and Twitter keeps increasing at a rapid pace amounting to 82 percentage of the world's online population. Social network usage has increased by 64% since 2005.. Proposed work to use non verbal behavior to identify and prevent multiple accounts in social media. Thereby security and privacy issues will be decreased with accuracy. The main contributions of this work can be propose a computationally efficient method (applicable to all social media classifications) for detecting identity deception through the use of non-verbal user activity in the social media environment.

Index Terms: Algorithm; Deception; Identity, Performance; Social Media

I. INTRODUCTION

Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse complex and of a massive scale. Big data with using concepts of hadoop and Mapreduce.Big Data is all deals finding a needle of value in a haystack of unstructured information[1]. Big data regularly includes data sets with sizes beyond the ability of commonly used software tools to capture and manage and process data within a tolerable elapsed time. Planning a big Data architecture is not about understanding just what is different[5]. It's also deals how to integrate what's new and what you already have – from database-and-BI infrastructure to IT tools and end user applications [3]. Big data specifies to large datasets that are difficult to store, search, share, visualize and analyse. Big Data is sized in peta, hexa and zeta bytes. it's not just about volume the approach to analysis contends with data content and structure that cannot be anticipated [6]. These analytics and the science behind them filter low value or low-density data to reveal high value or high-density data.

1.1 Big Data Characteristics

Volume- It refers to the amount of data. it supports high volume of data even terabytes and beta bytes and so on. The quantity of data that is generated is very important in this context [4]. It defines the size of the data which determines the value and potential of the data.

Variety - It refers to the variety of data. it supports many types of data. For example image, audio, video etc. It helps the people who are closely analysing the data and are associated with it to effectively use the data to their advantage and thus upholding the importance of the Big Data [9].

Velocity – It refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges [7].

Variability – It refers to the process of being able to handle and manage the data effectively.

Veracity – It refers the quality of the data being captured can vary greatly. Exact analysis depends on the veracity of the source data [2]. Veracity refers to the level of quality and trustworthiness that can be ascribed to a data set.

1.2 Key Areas That Big Data Analytics May Influence Are Detailed Below

Data management — There are potential savings in time and money if agencies implemented smarter data management practices that were conscious of the needs of big data analysis.. For example through better business process management, redundant data collection processes can be reduced by reusing data collected from separate processes [8].

Personalisation of services—Big data analytics may produce value by revealing a clear picture of an individual customer or customer group. Big data is able to achieve this due to its characteristic granularity. This granularity may assist in unlocking the possibility of personalised services tailored to the individual and delivered by government. Problem solving and predictive analytics the unification of multiple datasets from disparate sources in combination with advanced analytics techniques and technologies will advance problem solving capabilities and in turn will improve the ability of predictive analytics to reveal insights that can effectively support decision-making.

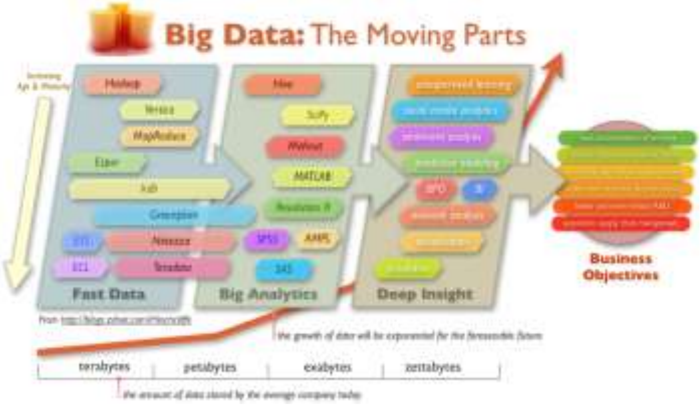


Fig 1.1 Big Data Technology

Productivity and efficiency — The analysis of big data sources can be used to identify cost savings and opportunities to increase efficiency. it will directly contribute to an improvement in productivity. It can in turn help to encourage further innovation.

1.3 What the Future Looks Like

A successful big data strategy is expected to assist in realising each of the priority areas observed in the ICT Strategy. The delivery of better services — big data analytics will allow government agencies to deliver more personalised services that are tailored to meet citizen's needs and preferences. For example the identification of individuals or groups who are eligible for certain entitlements without the need for them to be aware of or explicitly apply for that benefit. Improved efficiency of government operations — more effective use of big data for predictive analysis will allow government agencies to better assess risk and feasibility and detect fraud and error. Open engagement —These engagements will help to build knowledge, spark ideas, generate growth and better inform decisions and solutions that meet the needs of the government, both on a national and local level.

1.4 Challenges

Meeting the challenges presented by big data will be difficult. The volume of data is already enormous and increasing every day. The velocity of its generation and growth is increasing, driven in part by the proliferation of internet connected devices. Current technology, architecture, management and analysis approaches are unable to cope with the flood of data, and organisations will need to change the way they think about, plan, govern, manage, process and report on data to realise the potential of big data.

1.4.1 Privacy, Security and Trust

Big data sources, the transport and delivery systems within and across agencies, and the end points for this data will all become targets of interest for hackers, both local and international and will need to be protected [11]. The potential value of big data is a function of the number of relevant, disparate datasets that can be linked and analysed to reveal new patterns, trends and insights.

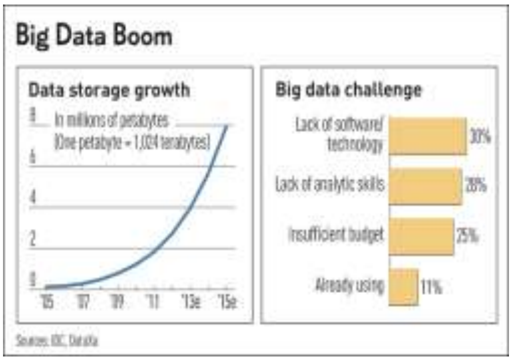


Fig 1.2 Challenges Handled in Big Data

1.4.2 Data Management and Sharing

The processes surrounding the way data is collected, handled, utilised and managed by agencies will need to be aligned with all relevant legislative and regulatory instruments with a focus on making the data available for analysis in a lawful, controlled and meaningful way.

II. HADOOP

IT defined as a framework for running applications on large clusters of commodity hardware which produces huge data (petabytes – zetabytes) and to process it.Open source Apache Software Foundation Project.

2.1 Hadoop Includes

HDFS a distributed file system to distribute data.A File System on multiple machines which sits on native file system .It supports processing incase of any hardware Failure due to usage of Commodity machines, failure is a common phenomenon and designed for failure also it supports simple Coherency Model. it used to write Once and read Many Times. Map/Reduce HDFS implements this programming model. It is an offline computing engine. Handles distributed Applications.

Hadoop Daemons:

1. Name Node- Stores the metadata that means information about the files and blocks.
2. Data Node -Serve read/write requests from clients and Perform replication tasks upon instruction by name node.
3. Secondary Name Node- Copies Fs Image and Transaction Log from Name Node to a temporary directory. Merges FS Image and Transaction Log into a new FS Image in temporary directory.

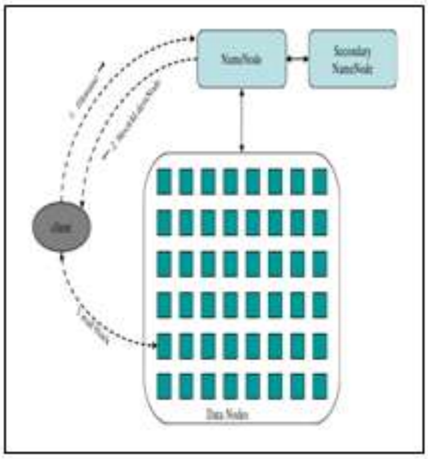


Fig 2.1 HDFS Architecture

1. Job Tracker -Accepts MR jobs submitted by users.Assigns Map and Reduce tasks to Task trackers. Monitors task and task tracker status, re executes tasks upon failure.
2. Task Tracker -Run Map and Reduce tasks upon instruction from the Job tracker. Manage storage and transmission of intermediate output**.**

2.2 Characteristics

Scalable is that new nodes can be added as needed and added without needing to change data formats, how data is loaded, how jobs are written, or the application on top. Cost effective is massively parallel computing to commodity servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it affordable to model all the data [13]. Flexible is schema-less and can absorb any type of data, structured or not from any number of sources. Data from multiple sources can be joined and aggregated in arbitrary ways enabling deeper analyses than any one system provide [12]. Fault tolerant is the system redirects work to another location of the data and continues processing. It defines high level abstracted framework for distributed processing of large datasets. it supports Fault Tolerant and Parallelization. Computation consists of two phase Map and Reduce.

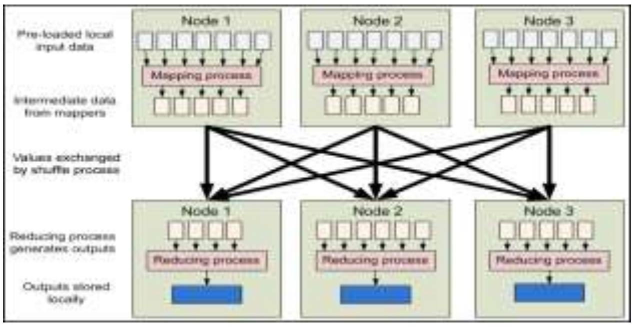


Fig 2.2 Map Reduce Pipeline

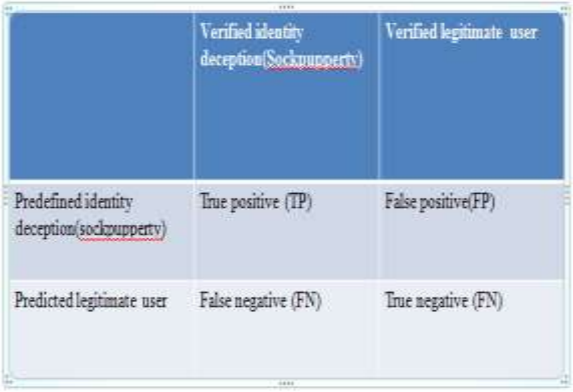
It is master-slaves architecture and computations occurs in multiple slave nodes and it tries to provide data locality as much as possible.

III. PROPOSED SYSTEM ARCHITECTURE

This chapter deals with the details of the proposed system. Identity deception is an important issue in the social media environment. The blocked users initiating new accounts, often called sock puppetry is widely known and existing efforts, it have attempted to detect such type of users, have been initially based on verbal behaviour. Although these methods yield a high detection rate of accuracy, they are computationally inefficient for the social media environment, its involving databases with large volumes of data. These past methods have mainly focused on detecting deception through verbal communication (e.g., voice or text). Proposed work to use non verbal behaviour to identify and prevent multiple accounts in social media. Issues of security and privacy will be decreased with accuracy. The main contributions of this work can be summarized as follows: propose a computationally efficient method (applicable to all social media classifications) for detecting identity deception through the use of non-verbal user activity in the social media environment.

3.1 Classification Matrix

The classification matrix is used to find the four metrics those are True positive (TP), True negative (TN), False positive (FP) and False negative (FN).



	Verified identity deception(Sockpupperty)	Verified legitimate user
Predefined identity deception(sockpupperty)	True positive (TP)	False positive(FP)
Predicted legitimate user	False negative (FN)	True negative (FN)

Table 3.1 Classification Matrix

3.2 System Architecture

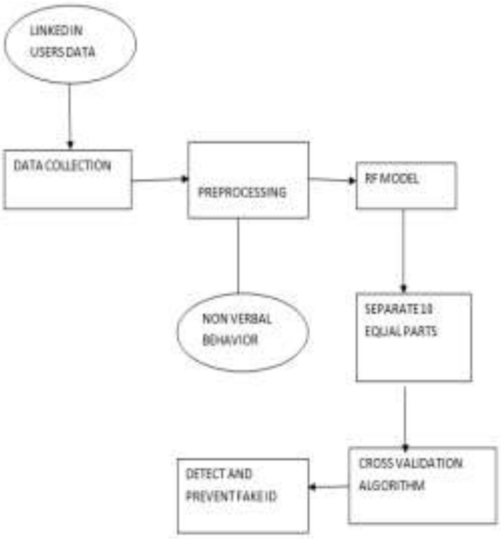


Fig 3.2 System Architecture

The Ten Times Ten Fold Cross Validation Algorithm it contains the following steps:

Step 1: Set a predefined number w

Step 2: Set random seed S= w*n*10

Step 3: Build RF model

Step 4: Classification matrix

Step 5: Calculate Recall, Precision and F-Measure.

Using this matrix, derive results to measure the following performance metrics in order to evaluate the performance of models for proposed method**:** recall (the fraction of valid sock puppet cases that are returned), precision (the amount of returned cases that are valid sock puppet cases), F-measure (the test of a model's accuracy bounded between0 and 1 that combines recall and precision), accuracy (the fraction of true positives and true negatives returned over the total number of cases), false positive rate (indicating the rate of falsely identified sockpuppets), and Matthews Correlation Coefficient (MCC) (a performance metric used in machine learning that provides a balanced result even if cases in the sample vary substantially in size).

These performance metrics are formally defined as follows

Recall= $\frac{TP}{TP+FN}$

Precision = $\frac{TP}{TP+FP}$

F-Measure=$\frac{2\ast Precision\ast Recall}{Precision+Recall}$

Accuracy =$\frac{TP+TN}{TP+TN+FP+FN}$

FPR=$\frac{FP}{FP+TN}$

MCC=$\frac{TP\ast TN-FP\ast FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

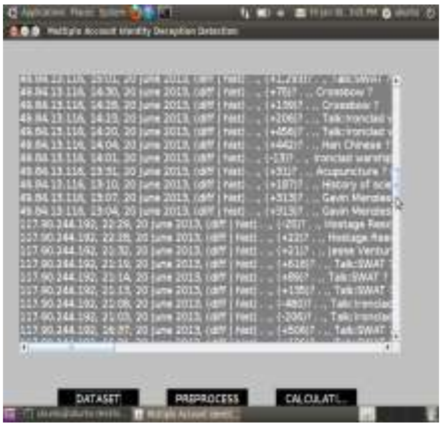
IV. RESULTS

4.1 Deception and Identity Deception

Deception has been defined as the deliberate transfer of false information to a recipient that is not aware that the information received has been falsified [6], [10]. Human deception is motivated by instrumental (goal-driven), relational (relationship-driven) and identity-driven goals [12]. Deception is achieved by manipulating content, the communication channel, the sender information, or any combinations of these three components [9]. Identity deception (a subcategory of deception) focuses on manipulating the sender's information [20] and can be divided into three categories: identity concealment (e.g., concealing or altering part of an individual's identity), identity theft (e.g., mimicking another person's real identity) and identity forgery (e.g., forging a fictional identity).

4.2 The Wikipedia Environment

Wikipedia is a free online encyclopedia in which everyone can contribute without an account (anonymously when only IP address is visible) and with an account using a pseudonym or even real name. Wikipedia operates on the concept of namespaces where each namespace is meant to include a specific type of content (or pages).

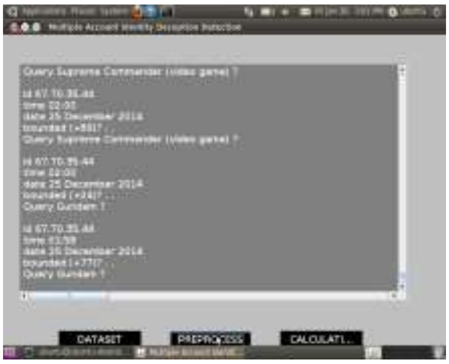


```
49.84.13.116, 15:01, 20 June 2013, (diff | hist) . . (+1,233)? . . Talk:SWAT
49.84.13.116, 14:30, 20 June 2013, (diff | hist) . . (+76)? . . Crossbow ?
49.84.13.116, 14:28, 20 June 2013, (diff | hist) . . (+139)? . . Crossbow ?
49.84.13.116, 14:19, 20 June 2013, (diff | hist) . . (+206)? . . Talk:Ironclad w
49.84.13.116, 14:20, 20 June 2013, (diff | hist) . . (+456)? . . Talk:Ironclad w
49.84.13.116, 14:04, 20 June 2013, (diff | hist) . . (+442)? . . Han Chinese ?
49.84.13.116, 14:01, 20 June 2013, (diff | hist) . . (-13)? . . Ironclad warship
49.84.13.116, 13:51, 20 June 2013, (diff | hist) . . (+31)? . . Acupuncture ?
49.84.13.116, 13:10, 20 June 2013, (diff | hist) . . (+187)? . . History of sci
49.84.13.116, 13:07, 20 June 2013, (diff | hist) . . (+313)? . . Gavin Menzies
49.84.13.116, 13:04, 20 June 2013, (diff | hist) . . (+313)? . . Gavin Menzies
117.90.244.192, 22:29, 20 June 2013, (diff | hist) . . (-20)? . . Hostage Resc
117.90.244.192, 22:28, 20 June 2013, (diff | hist) . . (+22)? . . Hostage Resc
117.90.244.192, 22:32, 20 June 2013, (diff | hist) . . (+21)? . . Jesse Ventur
117.90.244.192, 22:19, 20 June 2013, (diff | hist) . . (+618)? . . Talk:SWAT ?
117.90.244.192, 22:14, 20 June 2013, (diff | hist) . . (+89)? . . Talk:SWAT ?
117.90.244.192, 22:13, 20 June 2013, (diff | hist) . . (+135)? . . Talk:SWAT
117.90.244.192, 22:08, 20 June 2013, (diff | hist) . . (-480)? . . Talk:Ironclad
117.90.244.192, 21:03, 20 June 2013, (diff | hist) . . (-204)? . . Talk:Ironcla
117.90.244.192, 20:37, 20 June 2013, (diff | hist) . . (+506)? . . Talk:SWAT
```

Fig 3.1 Dataset Collection

4.3 Deception Detection

Deception detection theories are divided into those that are based on leakage cues (cues sent by the deceiver unwillingly due to factors such as cognitive overload) and strategic decisions (cues indicative of deception that are willingly transmitted by a deceiver in order to ensure deception success). To detect deception, both categories pick up cues from verbal and non-verbal communications. Three of the most popular theories used in the deception field are Interpersonal Deception Theory (IDT), Leakage Theory (LT), and Expectancy Violations Theory (EVT) [21], [26]. In IDT, deception is seen as a series of exchanges between the deceiver and the victim.



```
Query Supreme Commander (video game) ?

id 67.70.35.44
time 02:03
date 25 December 2014
bounded (+83)? . .
Query Supreme Commander (video game) ?

id 67.70.35.44
time 02:00
date 25 December 2014
bounded (+24)? . .
Query Gundam ?

id 67.70.35.44
time 01:59
date 25 December 2014
bounded (+77)? . .
Query Gundam ?
```

Fig 3.2 Pre-processing

4.4 Contributions of This Work

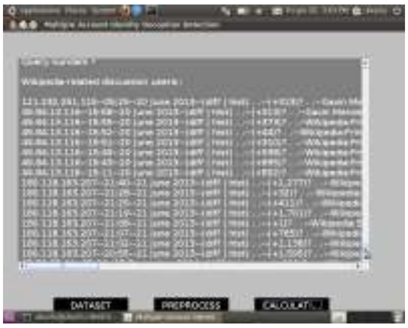
The main contributions of this work can be summarized as follows:

We propose a computationally efficient method (applicable to all social media classifications [1]) for detecting identity deception through the use of non-verbal user activity in the social media environment. This contribution ensures that a relatively high level of overall detection accuracy is obtained that is comparable to similar methods that make use of verbal communication [5], [6] but with lower computational overheads.

To demonstrate the computational efficiency (to withstand the immense traffic experienced by social media services) of our proposed non-verbal method to deception detection we use publicly available data from Wikipedia and machine learning algorithms. Finally, we present design guidelines for designers and developers interested in implementing this method as an added level of security for their social media communities and additional considerations based on various social media classifications in existence today.

4.5 Non-Verbal Behaviour Variables

We used simple and more complex variables to represent user behaviour. Variables of online non-verbal behaviour fall under two major categories: time-independent and time-dependent (henceforth these variables are denoted with index *t).*



Wikipedia-related discussion users :

Fig 3.3 Calculation

V. CONCLUSION

Identity deception has become an increasingly important issue in the social media environment using Sock puppet algorithm which have attempted to detect such users, have been initially based on verbal behaviour. These algorithms yield a high detection accuracy rate but they are computationally inefficient for the social media environment, it involves databases with large volumes of data. A detection method based on non-verbal behaviour for identity deception, it can be applied to multiple types of social media. The method gives high detection accuracy and being computationally efficient for the social media environment when compared to previous methods.

REFERENCES

[1] Anna Squicciarini,Christopher Griffi "An Informed Model of Personal Information Release in Social Networking Sites " arXiv:1206.0981v1[cs.SI] 5 June 2012.

[2] Anna Squicciarini,Sushama Karumanchi,Dan Lin,Nicole DeSisto "Identifying Hidden Social circles for Advanced Privacy Configuration "computer and security 41(2014) 40-51.

[3] T. Solorio, R. Hasan, and M. Mizan, "A case study of sockpuppet detection in wikipedia," in Proc. Workshop Lang. Anal. Social Media, 2013, pp. 59–68.

[4] Mauro Conti, Radha Poovendran, Marco Secchiero "FakeBook: Detecting Fake Profiles in On-line Social Networks" ACM International Conference on Advances in Social Networks Analysis and Mining IEEE 2012.

[5] Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis and Evangelos P. Markatos "Detecting Social Network Profile Cloning" Foundation for Research and Technology Hellas{kondax, polakis, sotiris, markatos}@ics.forth.gr.

[6] Racha Ajami, Nabeel Al Qirim, Noha Ramadan "Privacy Issues in Mobile Social Networks" Procedia Computer Science 10 (2012) 672 – 679.

[7] G. Alan Wang, Student Member, IEEE, Hsinchun Chen, Fellow, IEEE, Jennifer J. Xu, and Homa Atabakhsh "Automatically Detecting Criminal Identity Deception: An Adaptive Detection Algorithm" IEEE Transactions On Systems, Man, And Cybernetics—part A: Systems And Humans, Vol. 36, No. 5, September 2010.

[8] X. (Sherman) Shen, “Security and privacy in mobile social network [Editor’s Note],” IEEE Netw., vol. 27, no. 5, pp. 2–3, Sep./Oct. 2013.

[9] T. Solorio, R. Hasan, and M. Mizan, “A case study of sockpuppet detection in wikipedia,” in Proc. Workshop Lang. Anal. Social Media,2013, pp. 59–68.

[10] G. A. Wang, H. Chen, J. J. Xu, and H. Atabakhsh, “Automatically detecting criminal identity deception: An adaptive detection algorithm,”IEEE Trans. Syst., Man, Cybern. A, Syst. Humans, vol. 36, no. 5,pp. 988–999, Sep. 2012.

[11] C. Griffin and A. Squicciarini. Toward a game theoretic model of information release in social media with experimental results. In Proc.2nd Workshop on Semantic Computing and Security,San Francisco, CA,USA, May 28 2012.

[12] K. P. N. Puttaswamy and B. Y. Zhao, Preserving privacy in location-basedmobile social applications, in Hotmobile’10 Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications, Ohio, USA, 2010.