ISSN 2395-1621

# Identification of Anomalous User Behavior in Twitter

[#1]Anjali Changale, [#2]Anjali Ghutke, [#3]Sonali Takke, [#4]Shital Peddawad

P.E.S Modern College of Engineering, Pune, India

**ABSTRACT**

**The internet is one of the biggest blessing to man given by technology. Modern era people can't imagine the life without internet; everyone wants to connect with each other through internet via social media at all the time. A major part of modern world people use an online communication system, such as email and social media sites (e.g. Twitter, Facebook and LinkedIn) for entertainment and business. They generate lots of digital data for various users' activity; introduce a system to understand user communication behavior. In online communication system users' who have anomalous behavior is the potential threat to the society. Here we are proposing a system for identification of anomalous user behavior in Twitter. We are identifying the impact of user in twitter through their tweets, profile, followers and messages to know what they think. We propose a novel visual analysis system for detecting, summarizing and interpreting via the unsupervised learning model, visualizes the behavior of suspicious users in behavior-rich context through novel visualization design, contextual views.**

ARTICLE INFO

**Article History**

Received: 28th December 2016

Received in revised form :

28th December 2016

Accepted: 30th November 2016

**Published online :**

**4th January 2016**

*Keywords:* **Anomalous behavior, unsupervised learning model.**

## I. INTRODUCTION

The size of a database in twitter has increased rapidly day-to- day, Due to which the anomalies are also increasing. Anomaly detection is a problem of finding patterns in data that do not confirm to expected behavior. These nonconforming patterns are often called as anomalies. The real-life or interesting relevance of anomalies is a key feature of anomaly detection.

We identify users' communication behavior by considering following features:

1. Behavior features: This features identifying user's role based on their posting and re-posting behaviors.
2. Context features: These feature categories based on topical keywords, amount of special tags or symbols or sentimental scores.
3. Interaction features: This feature describe user based on their communication pattern. How users communicate with others and how others respond to user.
4. User profile features: In that we check user profile information. e.g., in twitter a user may change its screen name frequent to pretend to others.
5. Temporal Features. In this category, posting, replying, receiving, interval frequency entropy, measure the regularity of certain types of user behaviors.
6. Network Features. Features such as users in and out degrees in twitter provide egocentric measurements of the social network structure in different aspects.

## II. SYSTEM

We are designing a novel visual analyze system by unsupervised learning model and visualization technique to detect anomalous user behaviour.

STUDY:

 

Different social media sites have different uses, strengths and advantages. Twitter could be called a 'real-time social networking' site, a place for sharing information as it happens and for connecting with others in real-time, often resulting in lasting friendships and contacts. A lot of people communicate via social networking like twitter. In the unsupervised anomaly detection, we are given an input as a set of user data where it's unlabeled data (or noisy data). The goal is to identifying anomalous user behavior. We train noisy data and apply a traditional anomalous detection algorithm over the data.

In unsupervised machine learning anomalous detection has many advantages over supervised machine learning anomaly detection. The main advantage of unsupervised machine learning that they don't need a labeled data. The unsupervised anomaly detection algorithm does not need train dataset. In addition, unsupervised anomaly detection algorithm can use to analyze historical data.

The major advantage of our system, it is flexibility. We can apply this system for different kind of data (e.g., email, Facebook, LinkedIn). We worked on TLOF algorithms for detecting anomalies. This algorithm is very efficient and can deal with high dimensional data.

## III. BRIEF DESCRIPTION OF SYSTEM

WORKING OF SYSTEM:

1. Twitter:
Twitter is an online social networking site which enables people to share information to each other. Twitter user dataset are available and extracted using Twitter4J API.

2. Data collection:
To fetch a data from twitter we need access of twitter, access obtained by creating a twitter application. Whenever we create application we get four access keys from twitter.

They are :
- Consumer Key
- Consumer Secret
- OAuth Access Token
- OAuth Access Token Secret

By using this key in Java program we are able to collect user data.

3. Flume:
Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

4. Pre-processing:
- Language check: Check whether the data is in English or not. If it is not in English then remove this part of data from data set.
- Pre-processing: Lemmatization or lemmatization in linguistics is the process of grouping together the different inflected forms of a word so they can use as a single item. In computational linguistics, lemmatization is the algorithm of determining the lemma for a given word. The process may involve various complex tasks such as understanding context and determine the part of speech of a word in a sentence required(E.g. knowledge of the grammar of a language).
- Remove Stop words: In computation, stop words are words which filter out before or after processing of natural language text(data). The stop words usually refer to the most common words in a language.
- Remove hash tags.
- Feature Extraction: Determine the right features for detecting malicious user behaviors corresponding to tweets collected from twitter.

5. Analysis:
Anomaly score is computed for User Behavior. By using TLOF algorithm anomaly score is computed for the user. We uses the time-adaptive local outlier factor model (TLOF) to identify anomalies as the sudden changes of user behaviors based on a set of features extracted for each user from the online communication data.

6. Visualization:
With the help of Jfree chart graphical representation of user behavior is displayed based on the anomaly score .

## IV. FUTURE SCOPE

The future of this data analysis field is vast.
1. It will use in different companies and organizations which use social media analysis to understand users behavior.
2. Text Analytic: It is the process of deriving the high quality information from the raw data such as unstructured data and predicting the analysis.
3. It can used in fraud detection and cyber intrusions.
4. Social Media data: It can use to mine email, Facebook and other social media Conversations for real-time decisions and identifying user nature.

## V. CHALLENGES

The system works on real-time data.
1. Capture and display the communication process through a simple and integrated visual design to help efficient visual comparison.

2. Capture how the activity patterns (e.g., how a user posts on Twitter),temporal patterns (e.g., frequency and duration of the communication process) and content patterns (e.g., the topics around which the interaction occurred) are important for revealing the insight of a user's behavior.

3. Design a generalized visualization to support anomaly detection of users based on various data collected from twitter.

## VI. CONCLUSION

The main aim of this paper is detection of anomalous user behavior on twitter via novel visual analysis system. The

 

data collection process will introduce us to the Java Twitter4J API. This project will help us to gain knowledge about installation and configuration of the Hadoop distributed file system. Among the many fields of analysis, there is one field where humans have dominated the machines more than any ability to analyze our impact in twitter.

## VII.ACKNOWLEDMENT

A special thanks to Prof. Mrs. J. M. Kanase for their guide and support.

## REFERENCES

[1]N. Cao, L. Lu, Y.-R. Lin, F. Wang, and Z. Wen. Social helix:visual analysis of sentiment divergence in social media. Journal of Visualization,2016.

[2]M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof:identifying density-based local outliers. In ACM sigmoid record, volume 29, pages 93–104. ACM, 2000

[3]Hadoop: The Definitive Guide Book by John White.

[4] tons of stuff, including a bunch of online papers

 