**VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY, PUNE**

**COMPUTER ENGINEERING DEPARTMENT**

**APRIL-MAY 2018**

# Synopsis


## Group number: PRJ/18-19/035

Group Members :

1.Shubhankar Bendre

2. Yogeshwar Birangal

3.Mahesh Mohite

4. Ganesh Udge

Email-ID : mahesh.mohite@viit.ac.in

Mobile no : +91 99238-35698

# Title : Statistical Analysis for Twitter Spam Detection

## Objective :

- To categorize Spam and Non-Spam tweets
- To categorize the tag based and link based tweets
- to work on performance evaluation measures like precision,recall,F-measure

## Abstract :

Twitter is one of the most widely used online social networks(OSNs) which has impacted significantly on media and users for socialising and communicating. The issue of security and unwanted data being gathered is paramount in every organisation. Therefore, we intend to get rid of bulk of unwanted or irrelevant tweets. This requires a user to initially be logged into the twitter account. Here we will have implemented the classification and decision tree algorithms. Naive Bayes for classification and Random Forest as the learning algorithm on the training dataset. Pre-labelling of the training dataset will be practised for segregation of spam and non-spam tweets. Later using it with the feature extraction mechanism and the main machine learning algorithm. A classification model will be used further in the project and used on the timely tweets.

From this project, we hope to build a web application for twitter platform for elimination of irrelevant tweets and ignoring the spam tweets.

## Briefs about Contents:

**Introduction :**

Online social networks (OSNs) like Twitter, Facebook and LinkedIn impacted significant on media and users for socializing and communication.Discussing Twitter for example, we can send/receive tweets and messages with images, videos, text and follow others we care about. Monthly active users - 336 million (as of first half of 2018). However, with the
growing number of Twitter users, Spam activities are increasing. Spam in twitter is usually referred as tweets that contain ads, that may redirect users to links that include phishing or malware download, etc. Twitter Spam leads to threatening of the cyberspace security. Therefore, Spam alertness and a need against Spam

propagation is necessary. Twitter has applied rules to regulate user usage behaviors such as restricting users from sending content, to cite other users repeatedly or just submit URLs.

**Technical Details :**

**Hardware Requirements:**

Processor - Pentium –III
Speed - 1.1 GHz
RAM - 256 MB(min)
Hard Disk - 20 GB
Floppy Drive - 1.44 MB
Key Board - Standard Windows Keyboard
Mouse - Two or Three Button Mouse
Monitor - SVGA

**Software Requirements:**

Operating System - Windows

Application Server - Apache Tomcat

Front End - HTML, JDK 1.7, JSP

Scripts - JavaScript.

Server side Script - Java Server Pages.

Database - My SQL 5.0

IDE - Eclipse

**Working :**

**Applications:**

- Use of this system is in online social networking for spam detection.
- Feature of spam tweets seems to be time varying.
- It is unable to detect the categorization of tweets on the basis of their types.

## References/Bibliography:

[1] H. Costa, F. Benevenuto, and L. H. C. Merschmann, “Detecting tip spam in location-based social networks,” in *Proc. 28th Annu. ACM Symp. Appl. Comput.*, 2013, pp. 724–729.

[2] S. Ghosh *et al.*, “Understanding and combating link farming in the Twitter social network,” in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 61–70.

[3] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, “Aiding the detection of fake accounts in large scale social online services,” in *Proc. Symp. Netw. Syst. Des. Implement. (NSDI)*, 2012, pp. 197–210.

[4] G. Stringhini, C. Kruegel, and G. Vigna, “Detecting spammers on social networks,” in *Proc. 26th Annu. Comput. Sec. Appl. Conf.*, 2010, pp. 1–9.

[5] K. Thomas, C. Grier, D. Song, and V. Paxson, “Suspended accounts in retrospect: An analysis of Twitter spam,” in *Proc. ACM SIGCOMM Conf. Internet Meas.*, 2011, pp. 243–258.