News Source Analysis Using Modelling and Classification of Tweets

Submitted in partial fulfillment of the requirements of the degree of
Bachelor of Engineering in Information Technology

By

Prachi Bhavsar (Roll Number:14101A0021)
Gautam Menon (Roll Number:14101A0050)
Vedhas Patkar (Roll Number:14101A0064)

Under the Guidance of
Prof. Varsha Bhosale
Department of Information Technology Vidyalankar Institute of Technology

Wadala(E), Mumbai 400 037

University of Mumbai

2017-18

CERTIFICATE OF APPROVAL

This is to Certify that the project entitled

News Source Analysis Using Modelling and Classification of Tweets

is a bonafide work of

Prachi Bhavsar (Roll number: 14101A0021)
Gautam Menon (Roll Number: 14101A0050)
Vedhas Patkar (Roll Number: 14101A0064)

submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of

Undergraduate in Information Technology

Signature of Guide Head of Department Principal

Project Report Approval for B.E.

This project entitled **News Source Analysis Using Modelling and Classification of Tweets** by

1. Prachi Bhavsar (Roll Number: 14101A0021)
2. Gautam Menon (Roll Number: 14101A0050)
3. Vedhas Patkar (Roll Number: 14101A0064)

is approved for the degree **Bachelor of Engineering in Information Technology**

Examiners

1. _____

2. _____

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Name of Student	Roll no.	Signature
1. Prachi Bhavsar	14101A0021	_____
2. Gautam Menon	14101A0050	_____
3. Vedhas Patkar	14101A0064	_____

Date:

Acknowledgement

It is a matter of great pleasure and privilege to be able to present this project, ”News Source Analysis Using Modelling And Classification of Tweets”. This project would not have been possible without kind support of many individuals and we would like to extend our sincere thanks to all of them. We are highly indebted to Prof. Varsha Bhosale for her guidance and constant supervision as well as for providing all the necessary information regarding the project. We are thankful to Dr. Varsha Turkar, Dr. Meenakshi Arya, Prof. Indu Anoop, Prof. Deepali Nayak for approving our project and providing us the appropriate suggestions and giving us their precious time. Words are inadequate in expressing our thanks to our classmates for their inputs and cooperation in carrying out the project work. We would like to express gratitude towards our parents for their kind cooperation and encouragement, which helped us in completion of this project. Sincere thanks to Prof. Ajitkumar Khachane for letting us take up this project.

Abstract

Man continuously feeds on information and evolves using that information. In recent times, the source of information is not limited to newspapers or news channels. Social media websites have made information accessible at the touch of a button. However, a lot of information that is available is incorrect and at times, such incorrect information can prove to be hazardous. Our project aims to reduce the spread of such information by devising an ingenious method where people can look for themselves, if the information is likely to be credible or not.

In certain regions people can be riled up by a rumour, and thus, tumultuous times require shutting down of the entire Internet or sources of information in that region for a long time. Besides these scenarios, there are people who intentionally or otherwise post misleading pieces of medical, financial or legal advice, which leads to spreading of incorrect information with disastrous consequences.

By means of our project, we aim to reduce the instances of unfounded rumours or potentially misleading information going viral, by providing people a way by which they can find out if some piece of information that they come across on Twitter is credible or not.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Definition

With the growing use of internet and social media platforms for broadcasting news and enlightening masses, there are increased instances of rumours and misinformation spreading through them. During high impact events that is, events which generate huge amount of interest among people, people rely social media sites like Twitter for quick updates. When an event of a sizeable magnitude and impact occurs, thousands of tweets are posted every hour. Due to the large amount of content generated on Twitter, it is hard to distinguish between credible and non-credible content in tweets. Therefore, it is essential to verify the content of tweets and categorize the credibility of tweets so that people are not only well informed but also correctly informed. By means of our project, people will be able to figure out how credible the content of a tweet in real time.

The project is basically divided into two aspects: User classification and Tweet Content Classification. Firstly, the users are assigned a credibility value: high, neutral or low. Secondly, the tweet posted by the user is checked for its credibility and accordingly a label: High credibility, Neutral credibility, Low Credibility will

be assigned. Combining the credibility value of both the user and the content, a recent tweet by the user will be judged for its trueness and thus, the content level of that tweet will be classified as very high credibility, high credibility, neutral credibility, low credibility and very low credibility according to the following table.

User	Tweet	Content
H	H	VH
H	N	H
H	L	N
N	H	H
N	N	N
N	L	L
L	H	N
L	N	L
L	L	VL

Table 1.1: Content Credibility

1.2 Aims and Objectives

1. The primary objective of the project is to create a product that would enable classifying content of tweets as credible or not in real time.

2. To ensure our contribution to the society, we will keep the website accessible as long as it is financially feasible for us to do so and introduce it to as many people as possible.

3. Our personal objective in this project is to develop a cutting-edge algorithm that will classify content into various levels of credibility, a problem that many high profile technology companies like Facebook and Twitter are looking to solve, thus helping us publish our work in highly regarded journals or conferences.

1.3 Scope

The project after completion will analyse each news imparting tweet for its credibility by first checking the credibility of the user and then taking into consideration the content of the tweet based on the source of the content, URL present in the tweet, length of the tweet, number of characters, etc. A credibility indicator based on these analyses will show the believability or credibility of the tweet thus providing the users with a credibility meter of the tweets.

1.4 Terminologies Used

- **Twitter**

  Twitter is a microblogging service that allows its members to publish short status updates known as tweets. User accounts and their status updates are public by default, accessible by the public via Twitter's application program interfaces (APIs). The large number of users, low privacy expectations, and easy-to-use API have made Twitter a target of abuse, whether relatively benign in the form of spam and disruptive marketing tactics, or malicious in the form of links to malware and phishing schemes.

- **Tweets**

  Tweets are short messages (limited to 140 characters) posted to a Twitter account using a browser, a stand-alone application, an API, or SMS messages. Information associated with each tweet includes the time at which the update was created and the source by which the status appears to have been posted. Users on Twitter can subscribe to the tweets of another account by choosing to follow that account.

- **Machine Learning**

  Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible; example applications include email filtering, detection of network intruders or malicious insiders

working towards a data breach, optical character recognition (OCR), learning to rank, and computer vision.

- **API**

  An application program interface (API) is code that allows two software programs to communicate with each other. The API defines the correct way for a developer to write a program that requests services from an operating system (OS) or other application. APIs are implemented by function calls composed of verbs and nouns. The required syntax is described in the documentation of the application being called.

- **Classification**

  In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

- **Cloud Storage**

  Cloud storage is a simple and scalable way to store, access, and share data over the Internet. Cloud storage providers such as Amazon Web Services own and maintain the network-connected hardware and software, while you provision and use what you need via a web application. Using cloud storage eliminates the acquisition and management costs of buying and maintaining your own storage infrastructure, increases agility, provides global scale, and delivers ”anywhere, anytime” access to data.

- **Regularization**

  In simple terms, regularization is tuning or selecting the preferred level of model complexity so your models are better at predicting (generalizing). If you don’t do this your models may be too complex and overfit or too simple

and underfit, either way giving poor predictions.

If you least-squares fit a complex model to a small set of training data you will probably overfit, this is the most common situation. The optimal complexity of the model depends on the sort of process you are modelling and the quality of the data, so there is no a-priori correct complexity of a model.

To regularize you need 2 things:

1. A way of testing how good your models are at prediction, for example using cross-validation or a set of validation data (you can't use the fitting error for this).
2. A tuning parameter which lets you change the complexity or smoothness of the model, or a selection of models of differing complexity/smoothness.

Basically you adjust the complexity parameter (or change the model) and find the value which gives the best model predictions.

Note that the optimized regularization error will not be an accurate estimate of the overall prediction error so after regularization you will finally have to use an additional validation dataset or perform some additional statistical analysis to get an unbiased prediction error.

Chapter 2

Literature Review

1. **An in-depth characterisation of Bots and Humans on Twitter**
   Bots place more external URLs in their tweets. Humans receive more favourites per tweet. Bots generate larger amounts of tweets than their human counterparts do, while they rely far more heavily on retweeting existing content and redirecting users to external websites via URLs.

2. **Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?**
   Using sentiment analysis to mine tweets and obtain emotion. Human users were found to display more emotions on average than bots. When humans express positive sentiment, they tend to express stronger positive sentiment than bots.

3. **Broker Bots: Analysing automated activity during High Impact Events on Twitter**
   Checks if bots broker content, that is direct requests to a malicious URL or not.

4. **Credibility Ranking of Tweets during High Impact Events**

A ranking system for credibility of tweets which uses the number of characters and the number of unique characters present in the tweet to analyse credibility.

5. **Detecting and analysing automated activity on Twitter**

   This paper analysed activity on Twitter and found out that automated sources of tweets used scheduling software and automated services while organic accounts used Twitter interface and other non-automated services.

6. **The Rise of Social Bots**

   An algorithm called Bot or Not? as defined where, bots retweet more content but are retweeted less often than humans, and that their usernames are longer and produced content and engagement levels are lower.

Chapter 3

Timeline

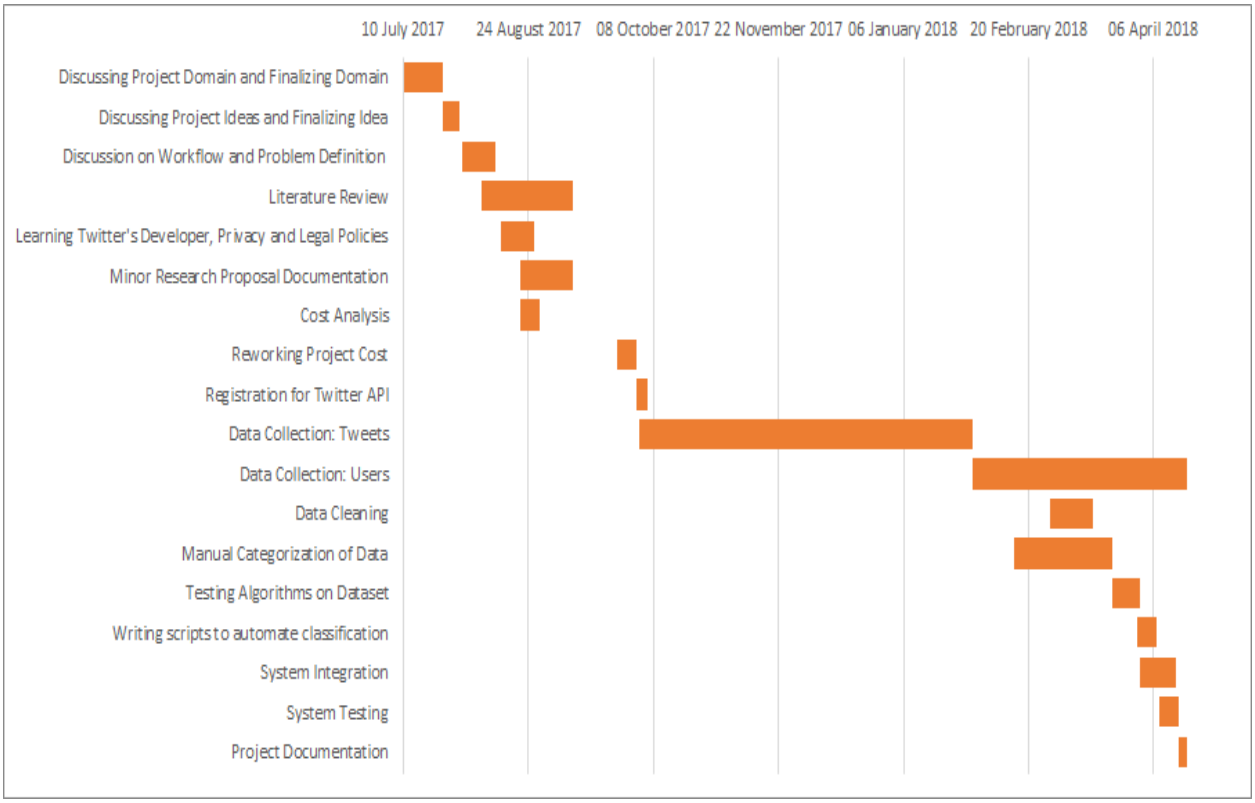


Figure 3.0.1: Timeline

Chapter 4

System Analysis

4.1 Process Model

Agile Model is based on the adaptive software development methods where there is no detailed planning and there is clarity on future tasks only in respect of what features need to be developed. There is feature driven development and the team adapts to the changing product requirements dynamically. The product is tested very frequently, through the release iterations, minimizing the risk of any major failures in future. Agile Model suits the project because there are three phases of development and deployment where a tweak or feature is added during every iteration of the SDLC phases.

Requirement Gathering and Analysis:

Requirements of the Credibility Ranking System are understood in detail. A study of existing systems that rank credibility is made, their drawbacks are understood. The parameters required to construct the model and data sets which can be used for testing the model are finalised. The requirements of both system and software to run the system and create the system is finalized

through extensive research.

Design:

Database schemas are finalised. Different software modules and algorithms are finalized. System interface is decided.

Construction:

The code for the system is written.

Testing:

In the testing phase, every single unit that is: prediction, user interface are tested before final deployment. This is done to make sure that the final deployed system is as bug-free as possible and can handle a large number of requests after being deployed.

Maintenance:

Software is changed for the encountered errors, it is also changed to accommodate changes in its external environment.

4.2 Feasibility Study

1. Executive Summary:

This project will be a model to improve the quality of information posted on Twitter.

2. Technology Considerations:

The users of the final system will only need a good quality browser so that they can check the credibility of tweets and tweeters in real time. Therefore, a web browser that supports scripts and that can be accessed when needed is the only requirement from the end user. Developing a model locally on a computer is not possible as the number of computations that are required for building the model and real time analysis of credibility are too high for a normal laptop or desktop computer. Therefore, an external computational and data storage provider will be utilized.

3 .Economic Feasibility:

Implementing this system will require a decent amount of capital as the costs of renting an external provider for data storage and real time prediction increase exponentially as the amount of data grows. However, by our estimates, we can build a working model by self financing the project so that it can be used by hundreds of users at a single time.

	Item	Estimated Expenditure
1	Data storage services	INR 5000
2	Machine Learning service for real time prediction	INR 30000
3	Access to publishing avenues for research	INR 5000

Table 4.1: Estimated Expenditure

4. Legal Considerations:

Our research shows that such a project does not violate any terms of usage of any service provider that we shall be using.

Chapter 5

System Design

5.1 Data Flow Diagrams

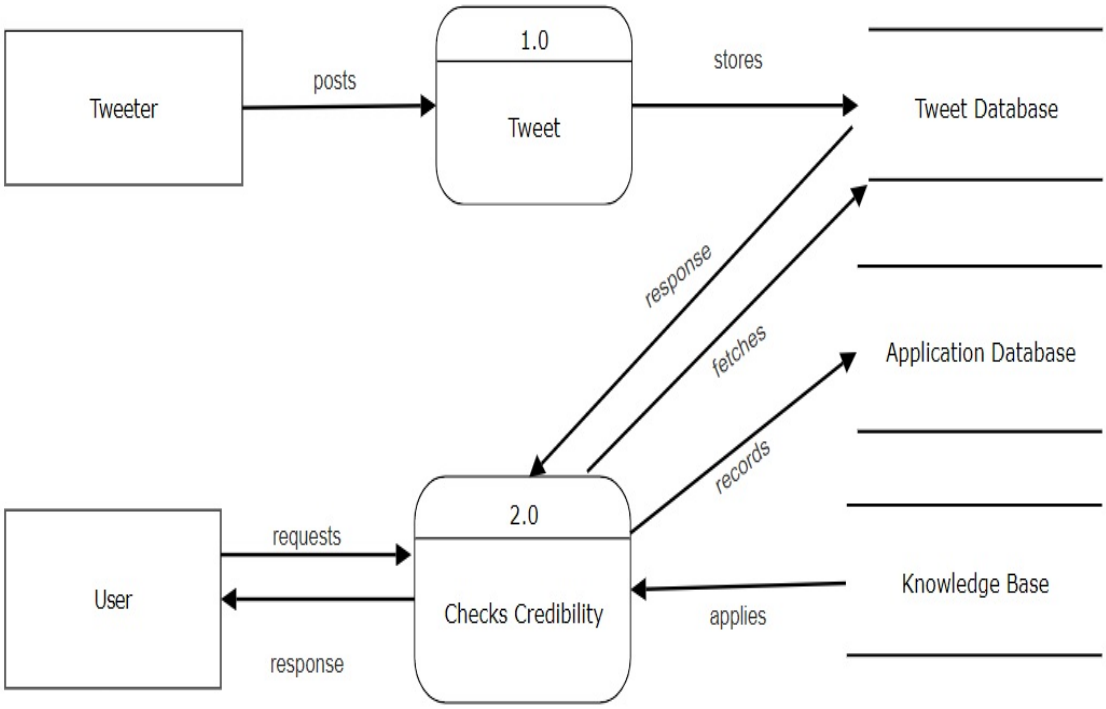


Figure 5.1.1: DFD Level 0

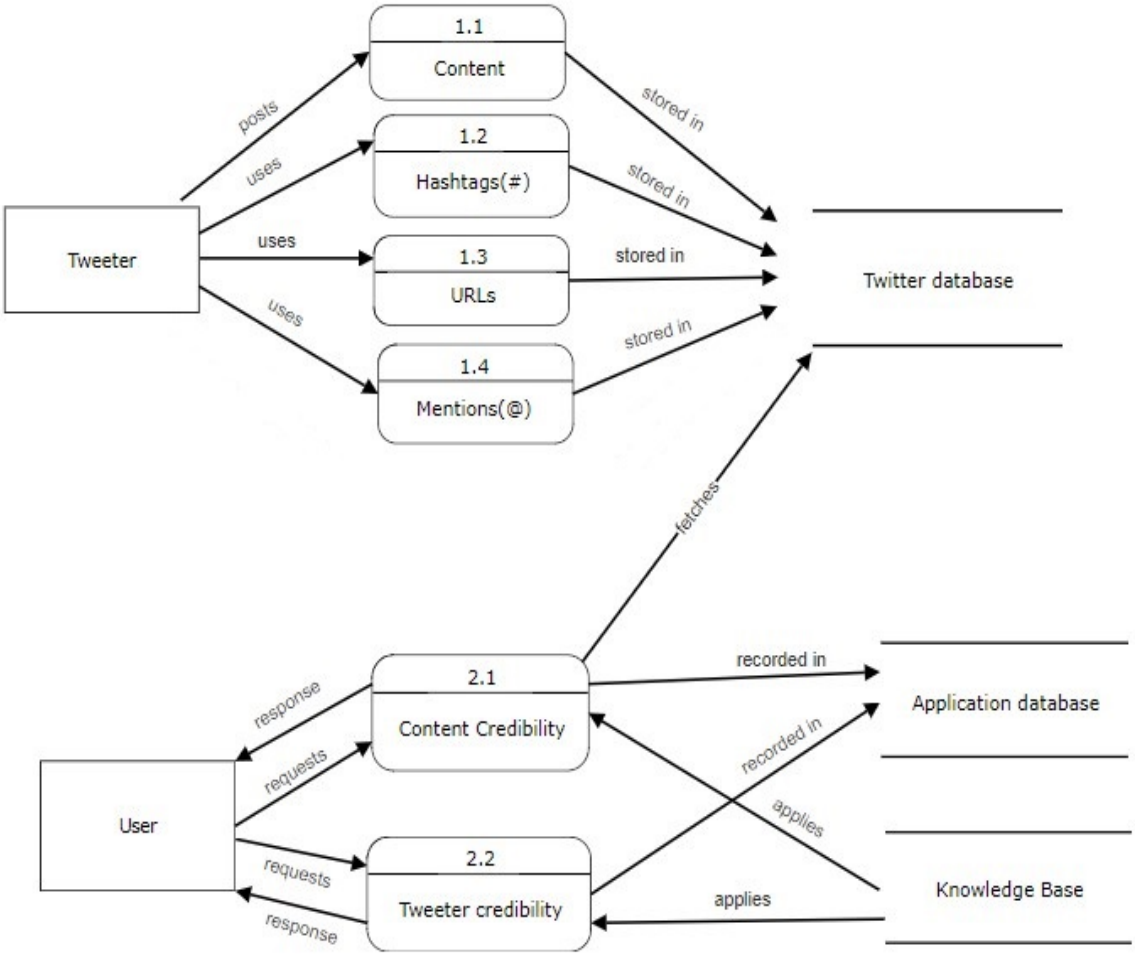


Figure 5.1.2: DFD Level 1

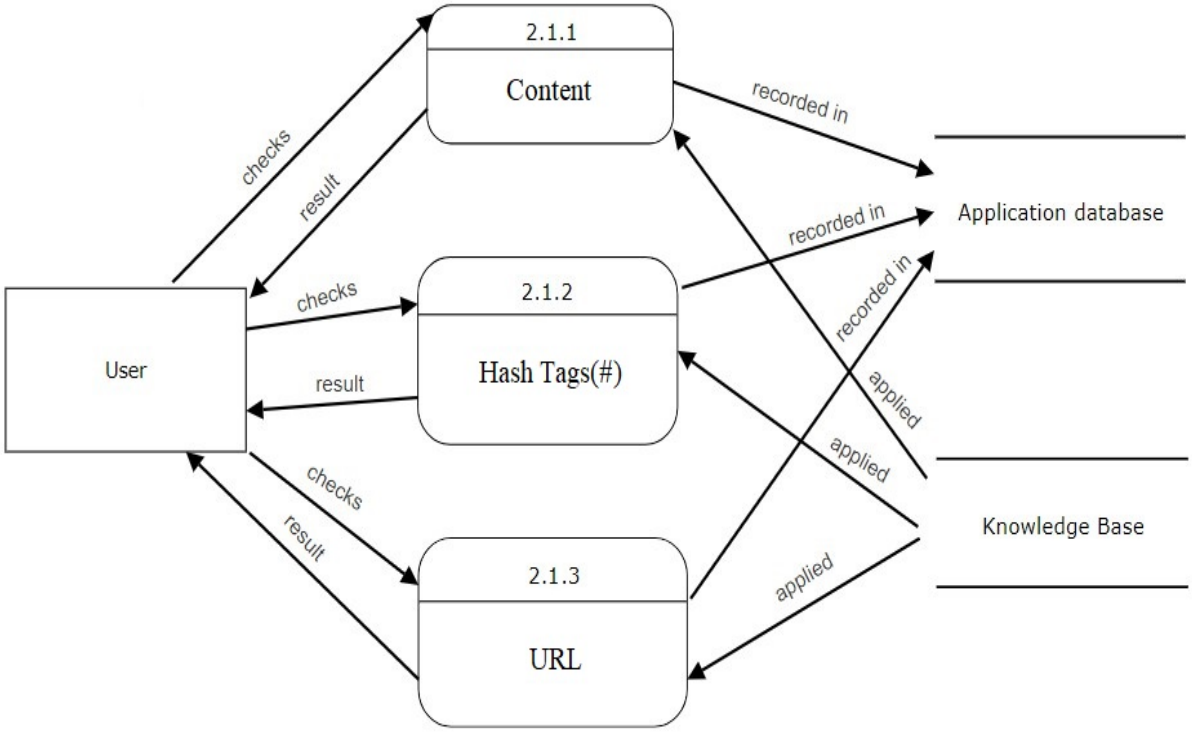


Figure 5.1.3: DFD Level 2

5.2 UML Diagrams

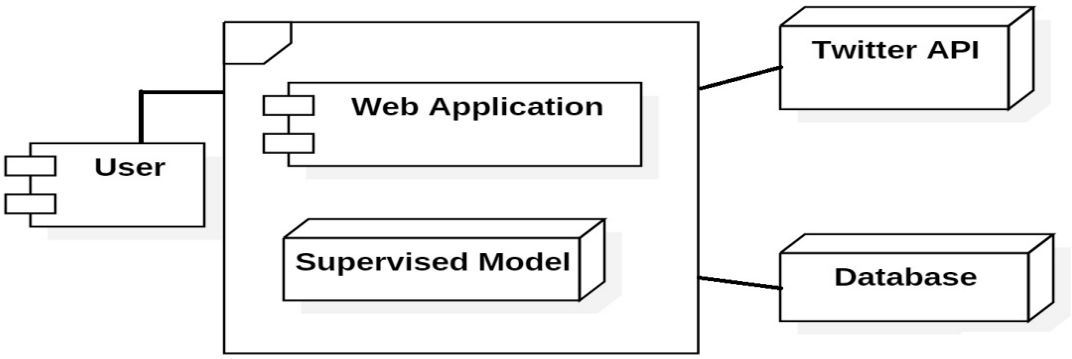


Figure 5.2.1: Deployment Diagram

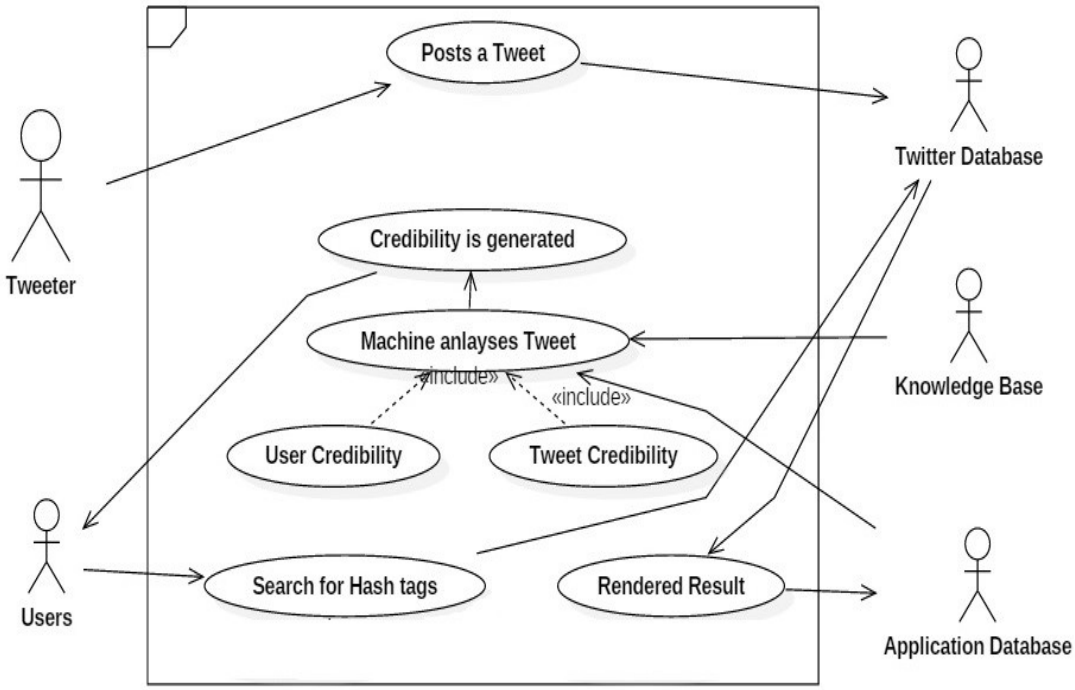


Figure 5.2.2: Use Case Diagram

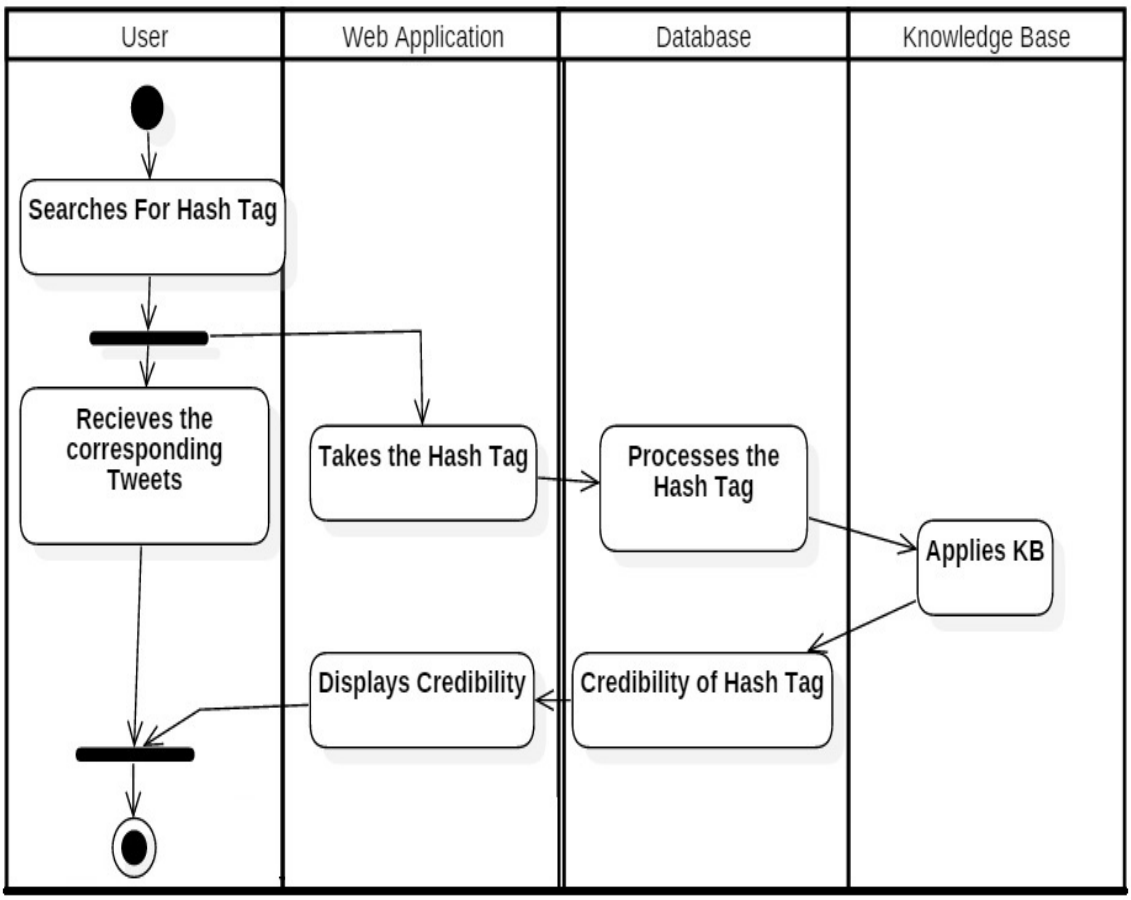


Figure 5.2.3: Activity Diagram

5.3 Table Structure

Data.csv:

	text	retweet_count	favorited	truncated	id_str
1	RT @OfficialNOI: Based on America's history of false flag operations...	328	FALSE	FALSE	983631818564943872
..	..	..	..	..	..

in_reply_to_screen_name	source	retweeted	created_at	in_reply_to_status_id_str	in_reply_to_user_id_str
NA	Twitter Lite"	FALSE	Tue Apr 10 09:04:35 +0000 2018	NA	NA
..	..	..	..	..	..

lang	listed_count	verified	location	user_id_str	description
"en"	67	FALSE	"SF Bay Area"	20886137	Neofolk
..	..	..	..	..	..

geo_enabled	user_created_at	statuses_count	followers_count	favourites_count	protected
TRUE	Sun Feb 15 01:49:11 +0000 2009	23717	846	121	FALSE
..	..	..	..	..	..

user_url	name	time_zome	user_lang	utc_offset	friends_count
http://tilhas.org	Michael Orion Powell	Alaska	"en"	-28800	2013
..	..	..	..	..	..

screen_name	country_code	country	place_type	full_name	place_id
mopowell	NA	NA	NA	NA	NA
..	..	..	..	..	..

place_lat	place_lon	lat	lon	expanded_url	url
NA	NA	NA	NA	NA	NA
..	..	..	..	..	..

Figure 5.3.1: JSON object obtained via Twitter API

ModelA.csv:

	text	id_str
1	RT @OfficialNOI: Based on America's history of false flag operations...	983631818564943872
..	..	..

ModelB.csv:

	source	listed_count	verified	user_created_at	statuses_count	followers_count	friends_count
1	Twitter Lite"	67	FALSE	Sun Feb 15 01:49:11 +0000 2009	23717	846	2013
..	..	..	..	..	..	..	..

ModelA_P.csv:

	text	id_str	Url_linked	Mentions	RT	No_ofwords	Hashtags	No_ofchars	Comma	Exclamations
1	RT @OfficialNOI: Based on America's history of false flag operations...	983631818564943872	TRUE	FALSE	FALSE	11-18	Zero	>90	FALSE	FALSE
..	..	..								

Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Credibility
1	0	0	1	0	0	0	0	10	NC
..	..	..	..	..	..	..	..	...	...

ModelB_P.csv:

	source	listed_count	verified	user_created_at	statuses_count	followers_count	friends_count	Ratio
1	Twitter Lite"	67	FALSE	Sun Feb 15 01:49:11 +0000 2009	23717	846	2013	0.420268
..	..	..	..	..	..	..	..	..

Score1	Score2	Score3	Score4	Score5	Score6	Score7	Credibility
1	2	0	1	0	2	2	NC
..	..	..	..	..	..	..	..

Figure 5.3.2: Model

5.4 Workflow

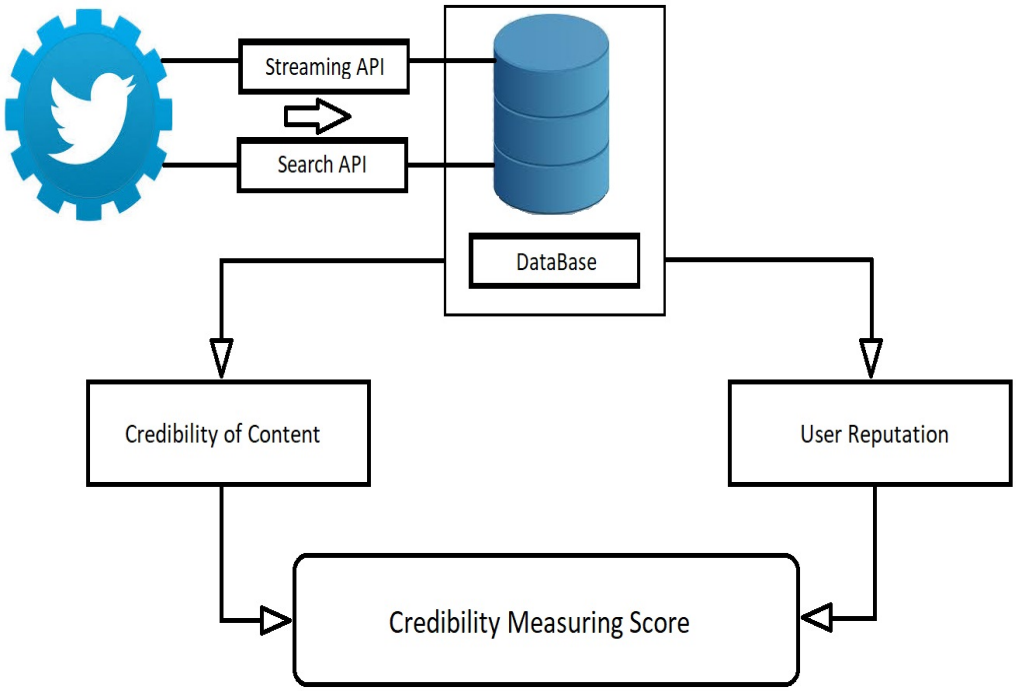


Figure 5.4.1: Project Flow

5.5 Technologies Used

- R Studio
- R Packages:ROAuth, streamR,rjson, RCurl, bitops
- Python 3.6
- Python packages: numpy, pandas, sci-kit,random
- Microsoft Excel
- Digital Ocean droplet
- Amazon Machine Learning
- Weka 3.8

Chapter 6

Implementation

6.1 Data Collection and Cleaning

6.1.1 Tweets

The Twitter API enables us to gather 42 parameters for a tweet. However, all parameters cannot be used for classification of tweets into categories as:

- The classes may be imbalanced or
- The parameters may not give any information

Therefore, we selected features based on past papers and our experience on Twitter as users. Initially, we chose these parameters for tweets:

- Tweet_ID: Identifier of the tweet (represented as T_ID)
- URL_Linked: Indicates if a tweet contains a hyperlink or not
- VIA: States if the content of the tweet was initially said by someone else

- RT: Indicates if a user is choosing to retweet someone else's tweet and quoting it
- Number of words in the tweet
- Number of hashtags in the tweet
- Number of characters
- Comma which states if a tweet contains a comma or not
- Exclamation which states if a tweet contains an exclamation mark or not
- Quotes which indicates if a tweet contains double quotes or not
- Mentions which indicates if that tweet contains a reference to another twitter user

However, we found out that, in our records, VIA and Quotes were hugely imbalanced and less the split between these values was more than 90% for a class label, i.e. the other class for that parameter, did not make up even 10% of the total values, and hence, we chose to remove these 2 parameters.

Developers who work on Twitter are bound by Twitter's Terms of Service and sharing user tweets which they may have used for their projects, is a violation of those terms and opens developers to legal action. As a result, we could scrape data for only two weeks at a time in the form of a JSON object. We then proceeded to manually classify 400 tweets which were picked by random sampling the following datasets:

- Tweets on Syria
- Tweets on 2016 US Presidential Election
- IPL

- Death of actress Sridevi
- Health News
- Narendra Modi
- Trending tweets

The data distribution was as follows:

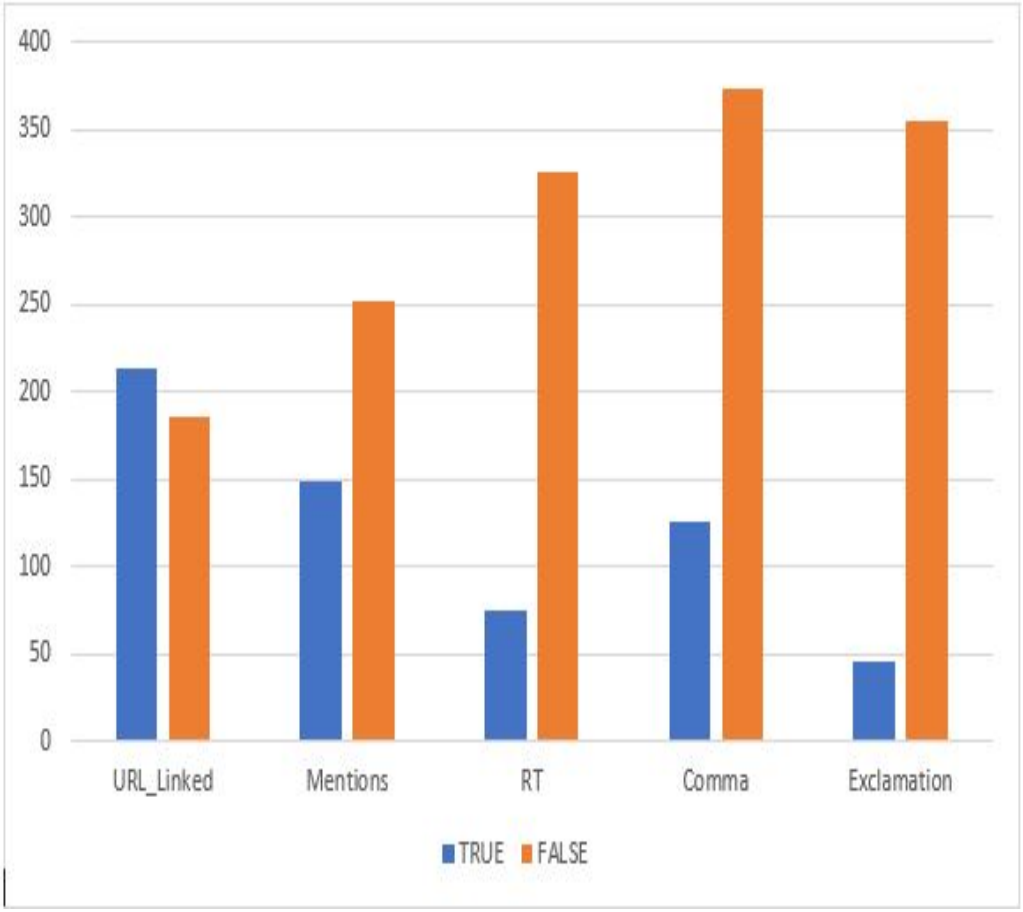


Figure 6.1.1: Data Distribution for Binary Values

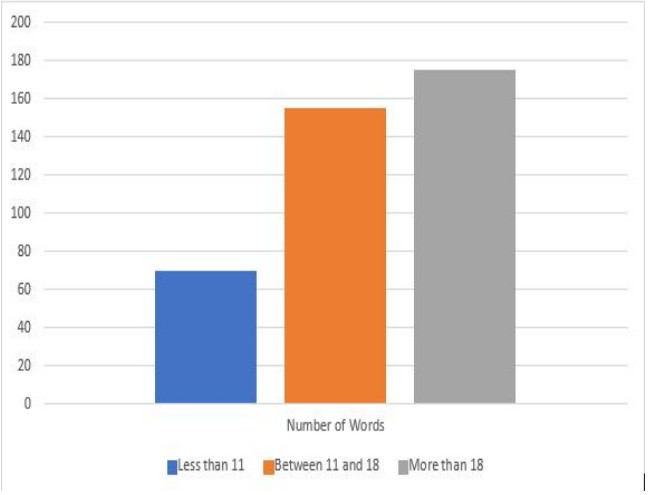


Figure 6.1.2: Data Distribution for Words

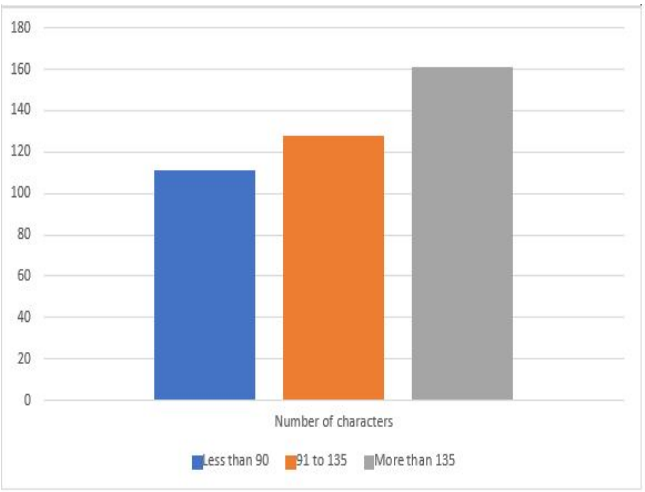


Figure 6.1.3: Data Distribution for Characters

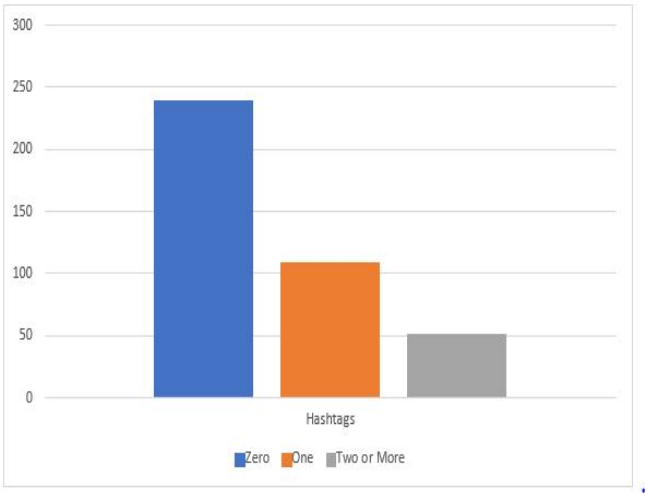


Figure 6.1.4: Data Distribution for Hashtags

As the problem of categorization was multiclass classification and not binary classification, we chose the following 3 algorithms to help us in building a model.

1. Multinomial Logistical Regression
2. Sequential Minimal Optimization
3. Random Forest

For creating a model, we used WEKA, and 3 methods of testing: Complete Test Data, 10-fold cross validation and Train-Test split of 70% and 80%. We found out that though Random Forest had the highest accuracy when all these testing methods were combined, multinomial logistical regression had took the least time.

6.1.2 Users

Based on our literature survey, we chose the following parameters for determining user credibility

- Source: The source from where the tweet was published
- Listed count: The number of public lists that a user is a part of
- Verified: Indicates if a user has been verified by Twitter
- Followers to friends ratio: The number of followers to friends for a Twitter user
- Account age: The age of an account in years
- Statuses count: The number of statuses posted by a user

For user categorization, we assigned each user a score as follows:

- Score 1: 1 if the source contained Twitter, 0 otherwise
- Score 2:

  High: if value between median and upper whisker of boxplot of all values or value within upper hinge of boxplot of outliers

  Moderate: if value between lower hinge to median of boxplot

  Low: too high or too low(outliers)
- Score 3: A combined score of Verified and Followers to Friends ratio
- Score 4: New if the account was made after 2014, Middle if between 2008 and 2013, Old otherwise
- Statuses count:

  High: if value between median and upper whisker of boxplot of all values or value within upper hinge of boxplot of outliers

  Moderate: if value between lower hinge to median of boxplot

  Low: too high or too low(outliers)
- Followers/Friends:

  High: if value between median and upper whisker of boxplot of all values or value within upper hinge of boxplot of outliers

  Moderate: if value between lower hinge to median of boxplot

  Low: too high or too low(outliers)

  We analyzed roughly 500 users, and chose Multinomial Logistical Regression to implement user categorization in the project.

6.2 Building the final system

Establishing connection with Twitter API

AuthandHandshake.fun(

Establish Handshake with ”https://api.twitter.com/”

Enter your twitter credentials and authorize the application

On sucessful authentication R redirects you to browser, enter the 6 digit pin from browser on R Console

)

Enter the Keyword

Fetch and Parse Tweets on required Hashtag

FetchandParse.fun(Tags)(

Search for Tweets.

Parse tweets and create data frames and CSV.

)

Clean the Data and generate required paramters for Tweet Content Credibility Processing

cleanAfxn.fun(

Read the required data from csv

Check for Linked URL : Url_linked

Check for Mention : Mentions

Checks if it is a RT : RT

Checks number of Words : No_ofwords

Count number of Hahtags(” ”) : Hashtags

Count number of charecters : No_ofchars

Check presence of Comma : Comma

Check for Exclamation mark : Exclamations

Write the processed data in CSV and into data frame

)

Clean the Data and generate required paramters for Tweeter Credibility Processing

cleanBfxn.fun(

Read the required data from csv

Check if Twitter is Source of information : Score1

Check the Listed count as a parameter : Score2

Check if user is verified or not. : Score_v

Check when Date of creation of account : Score4

Check statues count : Score5

Check followers and friend count : Ratio & Score6

Check for credibility with verified account and Ratio : Score3

Check credibility on basis of Verified + Ratio, Statuses Count and Listed Count : Score7

Write the processed data in CSV and into data frame.

)

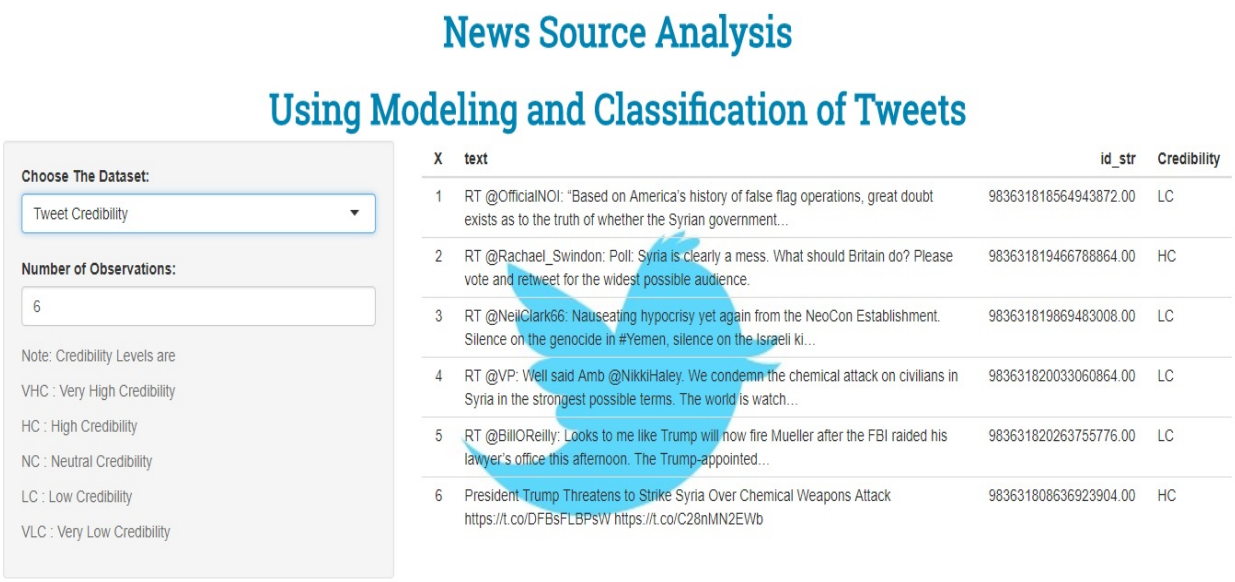


Figure 6.2.1: Tweet Credibility

Calculate Tweet Credibility.

modelAfxn.fun(

Read the required data from csv

Inputs to Train the machine.

Train the machine as ModelA_T

Stats of the Model

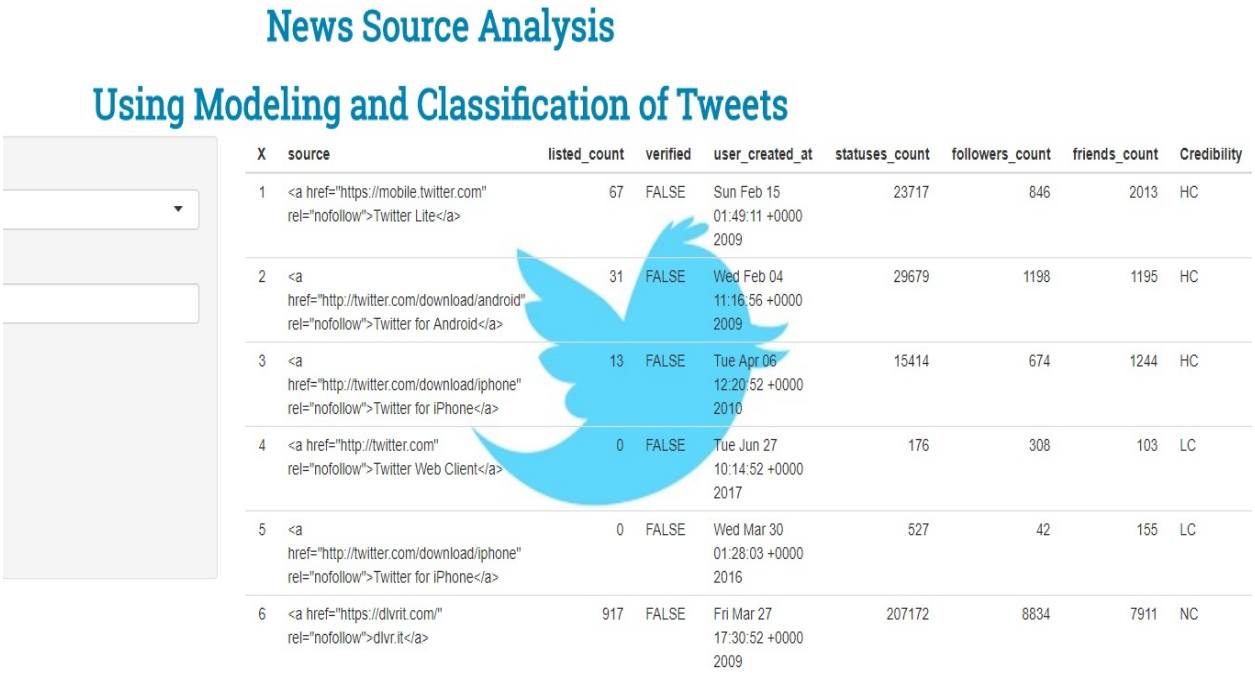
Run the model for new set.

Read the required data from csv

Get credibility for new Dataset.

Record the Credibility and write it into csv.

)



X	source	listed_count	verified	user_created_at	statuses_count	followers_count	friends_count	Credibility
1	Twitter Lite	67	FALSE	Sun Feb 15 01:49:11 +0000 2009	23717	846	2013	HC
2	Twitter for Android	31	FALSE	Wed Feb 04 11:16:56 +0000 2009	29679	1198	1195	HC
3	Twitter for iPhone	13	FALSE	Tue Apr 06 12:20:52 +0000 2010	15414	674	1244	HC
4	Twitter Web Client	0	FALSE	Tue Jun 27 10:14:52 +0000 2017	176	308	103	LC
5	Twitter for iPhone	0	FALSE	Wed Mar 30 01:28:03 +0000 2016	527	42	155	LC
6	dlvr.it	917	FALSE	Fri Mar 27 17:30:52 +0000 2009	207172	8834	7911	NC

Figure 6.2.2: Tweeter Credibility

Calculate Tweeter Credibility.

modelBfxn.fun(

Read the required data from csv

Inputs to Train the machine.

Train the machine as ModelB_T

Stats of the Model

Run the model for new set

. Read the required data from csv

Get credibility for new Dataset.

Record the Credibility and write it into csv.

)

```
Calculate Final Credibility.
  credibility.fxn(
    Final Credibility calculations.
    Write final scores into CSV
)

Output System as a R application

Define UI for application.
  ui <- fluidPage(
    Application title
    Sidebar with Inputs
    Sidebar Panel for inputs
    Input: Selector for choosing dataset
    Input: Number of obs
    Main panel for displaying outputs
)

Define Server Logic
  server <- function(input, output)
    Return the requested dataset.
    Show the first ”n” observations.

Run the application
  shinyApp(ui = ui, server = server)
```

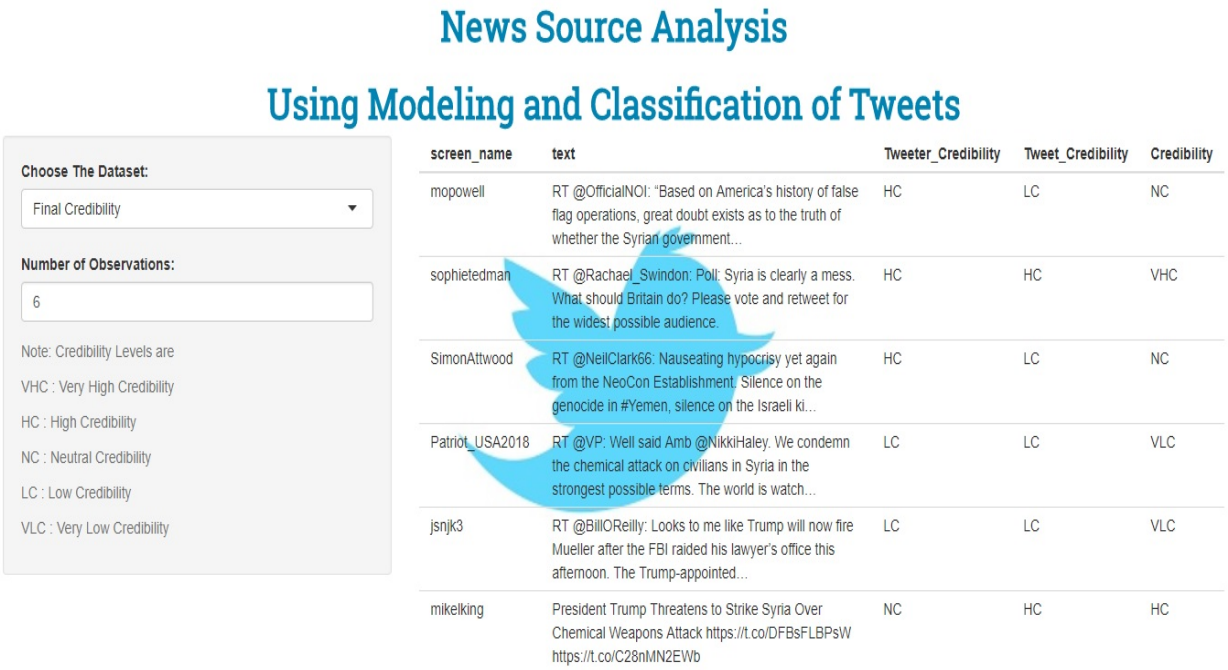


Figure 6.2.3: Final Output

Chapter 7

Conclusion

By working throughout the year, we think that we have created a very good content credibility system for Twitter which improves on previous models, and makes it easier for people to view the credibility of content that they encounter on Twitter. Overall, we feel that we have achieved our objectives of creating a high quality product, and aim to achieve publication in BDA 2018 at NIT Warangal in July.

References

[1] Machine Learning, Wikipedia.

[2] Application Programming Interface, Search Microservices.

[3] Statistical Classification, Wikipedia

[4] Cloud Storage, Amazon AWS

[5] A Quick Guide to Fake News Detection on Social Media, KDNuggets.

[6] Understanding Machine Learning Algorithms, KDNuggets.

[7] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, Jon Crowcroft, *"An in-depth characterisation of Bots and Humans on Twitter"*, arXiv, 2017.

[8] John P. Dickerson, Vadim Kagan, V.S. Subrahmanian, *"Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?"*, IEEE/ACM ICASMAM, China, 2014.

[9] Sudip Mittal, Ponnurangam Kumaraguru, *"Broker Bots: Analysing automated activity during High Impact Events on Twitter"*, arXiv, 2014.

[10] Aditi Gupta, Ponnurangam Kumaraguru, *"Credibility Ranking of Tweets during High Impact Events"*, Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, France, 2012.

[11] Chao Michael Zhang , Vern Paxson, *"Detecting and analysing automated activity on Twitter"*, ICPANM, 2011.

[12] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, Alessandro Flammini, *"The Rise of Social Bots"*, Communications of the ACM, 2016.

[13] SDLC - Agile Model, Tutorials Point.