



Inspiring Excellence

Evaluating user influence in Twitter based on hashtags using Data Mining

Supervisor: SuraiyaTairin

Submitted by:

MaheenAbsarNeha -13241007

Md. Intisher Rahman - 16341013

MahparaNuzhat - 14141003

SifatZereen - 13201012

Department of Computer Science and Engineering

School of Engineering and Computer Science

BRAC University

Submitted on: August 2017

Certificate

This is to certify that the work presented in this thesis entitled “Evaluating user influence on Twitter using Data Mining technique” is the outcome of the investigation carried out by us under the supervision of Lecturer Ms. Suraiya Tairin in the Department of Computer Science and Engineering, BRAC University, Dhaka. It is also declared that neither this thesis nor any part thereof has been submitted or is being currently submitted anywhere else for the award of any degree or diploma.

(Supervisor)

(Authors)

.....
Suraiya Tairin
Lecturer,
Department of Computer
Science and Engineering,
Brac University, Dhaka.

.....

Maheen Absar Neha
ID:13241007

.....

Mahpara Nuzhat
ID: 14141003

.....

Md. Intisher Rahman
ID: 16341013

.....

Sifat Zereen
ID: 13201012

Contents

Acknowledgements

Abstract

Introduction

1.1	Literature Review.....	5
1.2	Inferences Drawn out of Literature Review.....	5
1.3	Related Works.....	6
1.4	Motivation.....	7
1.5	Alternative approach to using STATISTICA.....	8

Preliminaries

2.1.1	Tweets and Retweets.....	9
2.1.2	Hashtags.....	10
2.1.3	Data Mining.....	11
2.1.4	Tweepy.....	12
2.2.1	Statistica.....	12
2.2.2	Text Mining.....	14
2.2.3	Typical Applications for Text Mining.....	16
2.2.4	Approaches to Text Mining.....	17
2.2.5	Issues and Considerations for "Numericizing" Text.....	18
2.2.6	Transforming Word Frequencies.....	19
2.2.7	Latent Semantic Indexing via Singular Value Decomposition.....	21
2.2.8	Incorporating Text Mining Results in Data Mining Projects.....	22

Design and Implementation

3.1	Implementation methodology.....	23
3.2	Data extraction.....	24

3.3	Statistica Installation.....	28
3.4	Importing data in Statistica.....	29
3.5	Frequency.....	29
3.6	Obtaining Graphical Representation of Tweets and Retweets.....	31
Result		
4.1	Extracted Tweets.....	33
4.2	Frequency of different hashtags.....	34
4.3	Frequency of hashtags for consecutive days.....	41
4.4	Comparison on retweets and other tweets.....	42
4.5	Generation of graphs to show different use of hashtags in different locations.....	44
Conclusion		
5.1	Conclusion.....	52
5.2	Limitations.....	52
5.3	Future Work.....	53
References		
6.1	List of References.....	54

List of Figures

1	A look at the STATISTICA graphical interface.....	13
2	How Twitter API operates.....	24
3	Twitter REST API Python code for extracting tweets by hashtag.....	25
4	Tweepy code for extracting tweets by location.....	26
5	How the csv file looked before formatting.....	27
6	After formatting the data for moving to hive table easily, the look of the csv file.....	28
7	Workbook containing tweets with hashtag '#antibullying' on 13-08-2017.....	31
8	Workbook containing tweets from few of the chosen dates.....	32
9	Workbook classifying different kinds of tweets.....	32
10	Tweets extracted without location filter, displaying the time zone.....	33
11	Tweets filtered by specified latitude/longitude: America's coordinates used.....	34
12	Frequency of tweets containing #antibullying hashtag on 13-08-2017.....	34
13	Frequency of tweets containing #antibullying hashtag on 14-08-2017.....	35
14	Frequency of tweets containing #stopbullying hashtag on 13-08-2017.....	35
15	Frequency of tweets containing #stopbullying hashtag on 14-08-2017.....	36
16	Frequency of tweets containing #harassment hashtag on 13-08-2017.....	36
17	Frequency of tweets containing #harassment hashtag on 14-08-2017.....	37
18	Frequency of tweets containing #cyberbullying hashtag on 05-08-2017.....	37
19	Frequency of tweets containing #cyberbullying hashtag on 07-08-2017.....	38
20	Frequency of tweets containing #xenophobia hashtag on 06-08-2017.....	38
21	Frequency of tweets containing #xenophobia hashtag on 07-08-2017.....	39
22	Frequency of tweets containing #racism hashtag on 05-08-2017.....	39
23	Frequency of tweets containing #racism hashtag on 07-08-2017.....	40
24	Frequency of tweets containing #sexism hashtag on 05-08-2017.....	40
25	Frequency of tweets containing #sexism hashtag on 07-08-2017.....	41
26	Frequency of tweets containing the hashtag '#harassment' from 13-08-2017 till 17-8-2017.....	41
27	Frequency of tweets containing the hashtag '#antibullying' from 13-08-2017 till 17-8-2017.....	42

28	Comparison of ‘#stopbullying’ tweets.....	43
29	Comparison of ‘#harassment’ tweets.....	43
30	Comparison of ‘#antibullying’ tweets.....	44
31	Comparison of ‘#cyberbullying’ tweets.....	44
32	Frequency of #antibullying across different locations.....	46
33	Frequency of #iambullied across different locations.....	46
34	Frequency of #cyberbullying across different locations.....	47
35	Frequency of #harrassment across different locations.....	47
36	Frequency of #racism across different locations.....	48
37	Frequency of #sexism across different locations.....	48
38	Frequency of #stopbullying across different locations.....	49
39	Frequency of #xenophobia across different locations.....	49
40	Frequency of all the hashtags with respect to all locations.....	50
41	The accuracy percentage of the location analysis for each hashtag.....	51

ACKNOWLEDGEMENTS

First and foremost, we would like to express our deep gratitude to our supervisor, Suraiya Tairin Pakhi, for introducing us to this wonderful and fascinating topic Big Data. We have learned from her how to carry on a research work, how to write, speak and present well. We thank her for her patience in reviewing our work, for correcting our proofs and language, suggesting new ways of thinking, leading to the right way, and encouraging us to continue our research work. We again express our indebtedness, sincere gratitude and profound respect to her for her continuous guidance, suggestions and wholehearted supervision throughout the progress of this work, without which this thesis would never have been possible.

We would also thank all the members of our research group for their valuable suggestions and continual encouragements. Our parents also supported us to the best of their ability. Our heartfelt gratitude goes to them.

Abstract

As more and more data is being processed and generated every day, it has become a tremendous challenge to process and analyze. Big data analysis is a process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information. It can help understand the information contained within data using specialized tools and applications for predictive analysis, data mining, text mining, forecasting and data optimization. We will be working with data from the most popular microblogging platform, Twitter, to study the social issues concerning various forms of harassment. Twitter users categorize status messages (Tweets) using hashtags, which are also used for searching specific topics or events. We can determine trends in Twitter-documented bullying among different demographics by analyzing the hashtags which represent different forms of social attacks, incidents of oppression, discrimination and cultural persecution. Out of the several tools used worldwide to interpret datasets, we will be using an advanced data mining tool called STATISTICA.

Chapter 1

Introduction

The internet has undergone a massive surge in social networks with enormous amounts of data being created and distributed every minute. Twitter is one of the most popular social media websites in the world. Twitter's speed and ease of publication have made it an important communication medium for people from all walks of life. The notion of community in this social networking world has also caught lots of attention. Studying Twitter is useful for understanding how people use new communication technologies to form social connections, maintain existing ones, spread useful information and bring about social change. Since its inception in March 2006, Twitter has reached over 310 million monthly active users and on an average over 500 million tweets are sent per day. "Tweets" are short messages with a maximum length of 140 characters. Twitter is useful because it is real time and information can reach a large number of users in little time. This makes it a substantially significant source of information for data mining.

Twitter enables users to create accounts with just their email addresses. Names (accurate ones), dates of birth, etc. are not required to become a member of the community. Unfortunately, this creates an anonymous environment where people feel like they can say anything they want without any repercussions. This includes attacking people for their race, gender, political affiliation and religion. Even if those tweets are reported and the accounts are terminated, a person can easily make another account almost immediately and repeat the same rude behavior. Twitter itself has taken measures to counter the high levels of abuse on the website. It has aimed to seize the creation of accounts of repeat offenders, show safer search results and prevent abusive tweets from being shown. More recently it has started putting online bullies in "time-out" when its algorithms identify tweets that appeared to be harassing – like ones which use abusive language.

It is also working to more efficiently penalize and expel accounts which violate its official "Twitter Rules." These policies bar users from directly or indirectly threatening, and abusing other users. Our study aims to illustrate how big of a problem online bullying has become in this age of online anonymity, by showcasing how many of the total tweets posted on Twitter throughout a day can

be construed as abusive or hateful. Another objective is to ascertain which hashtag used to bully people online has been used or posted about the most (e.g. sexism against racism). Users whose accounts were geo-tagged and whose tweets contained certain hashtags (like ‘#racism’) were also extracted and processed. In terms of geography, we also focused on countries like USA and India to determine where cyberbullying, which has become a big issue in recent years, was being written about the most. We will represent all the obtained data graphically by the use of graphs, pie charts, etc.

As Twitter provides an official API, it has been used by many to run studies on social media trends. However, Twitter does not require its users to provide information like age, gender and location which makes it difficult for researchers to accurately categorize users. Pennacchiotti and Popescu (2011) had to infer all this user information based on the “Bio” fields, where users could optionally provide a short (160-character length) personal descriptions. Most users did not write much of anything pertinent in that field so their study was not wholly accurate. They used GBDT (Gradient Based Decision Tree) to classify users and interpret their genders but that only worked on 8 out of 10 of users and the accuracy was very low. Studies like Go, Bhayani and Huang (2009) Pak and Paroubek (2010) sought to ascertain positive, negative and neutral tweets based on emoticons. Pak and Paroubek (2010) also searched for particular words to determine the mood of tweets, although they used a different method of doing so.

In this paper we have also referenced papers which directly address bullying trends on Twitter. Cortis (2015) identifies the most popular hashtags used on a fixed number of tweets. Zhu, Xu and Bellmore (2012) also provided people with a sentiment analysis report done on Twitter data. From these papers we deduced that if the sample size of the study is not large, it becomes less reliable. Also, despite growing numbers of Twitter data mining papers being published, a vastly accurate method of interpreting tweets have still missing and so is an accurate means of predicting trends on social media.

Data mining has become a very popular phenomenon in recent years. It is a process used by companies to turn raw data into useful information. It has greatly aided businesses by enabling them to predict customer behavior and patterns so that they can suitably produce and market their

products to get the greatest profits. Data mining depends on effective data collection and warehousing as well as computer processing.

Companies these days have to process very large amounts of data (in petabytes) and to do that efficiently, they use advanced analytics tools like SAS (Statistical Analysis System), R and STATISTICA.

Social media is now a reflection of our society: it lets people voice their thought and opinions at the click of a button and Twitter is one of the most popular ones out there. Twitter is social media website where people can post messages about anything within a 140 character limit. Users can also post photos, videos and links. These posts are known as “tweets”. Twitter users can gain followers (subscribers) and follow other accounts. They can also like,” retweet” and reply to tweets. A “retweet” is the act of sharing another user’s tweet.

In order to obtain information from Twitter, we used Twitter’s Search API, which is part of its REST API. The Search API is a REST service which enables developers to search for specific tweets in terms of the many parameters the API supports, like language, search term and location. We also used ‘Tweepy’, an open-sourced API, to determine how often terms like ‘cyberbullying’ and ‘harassment’ were talked about in particular countries, like United States of America, India, Canada, etc. We used Python programming language to write the scripts in both APIs, in order to extract the tweets.

We extracted a total of about 150,000 tweets and kept the data in CSV files, formatting them by using features like ‘Autofilter’ and ‘Text to Columns’. The platform we chose to process the data is STATISTICA, which is an analytic tool originally created by StatSoft. It offers many data mining features like Text Mining, Clustering and Partitioning. We divided the data by date and hashtag and imported them into STATISTICA, after which we generated graphs showing the frequency of tweets containing different hashtags in periods of a day and a few consecutive days. We also mined the tweets which contained the location of those tweets and derived trees and graphs through the use of text mining and SVD (Singular Value Decomposition). Additionally, we sorted tweets into

categories of 'Retweets' and 'Other Tweets' to determine how much of the content was original and how much it was not (retweets).

Previous Works and Motivation

1.1 Literature Review

Social networking services or sites (also known as Web 2.0 applications) such as Twitter, Facebook, Flickr and YouTube have revolutionized the way information is produced, shared and stored: anyone can provide information, access and comment on the information, as cited in the paper [22]. As Twitter does not record any information on user gender, age or location, Pennacchiotti and Popescu[3] stipulated these by looking at the 'Bio' fields of Twitter users. By streaming and documenting the tweets, ranking and classification can be done that is performing term search to group data by specific hashtags or words or phrases. GBDT (Gradient Based Decision Tree) was used for user classification and found the genders of about 80% users with very low accuracy. Go, Bhayani and Huang [2] used distant supervision to classify user sentiments from the tweets of the users by analyzing the emoticons used. In sentiment analysis, the happy emoticons were classified as positive sentiments, sad emoticons as negative sentiments and others as neutral. Pak and Paroubek [3] did linguistic analysis on a collected set of texts and categorized positive, negative and neutral emotions using both emoticons and certain words or adjectives. They detected superlative adjectives for happy sentiments and past adjectives and verbs such as 'lost', 'gone', 'missed', etc. for sad ones. To categorize the opinions and feelings of Twitter users about specific brands, Jansen, Zhang, Sobel & Chowdhury [4] did data mining by classifying tweets into six groups: 'no sentiment', 'wretched', 'bad', 'so-so', 'swell' and 'great'.

1.2 Inferences Drawn out of the Literature Review

It has been observed that the accuracy of term search extraction of data is not too high and still there have been a lot of study going to improve the pattern or trend analysis as the data spread over twitter is not always informative enough to predict the trends immediately. For increasing accuracy in sentiment analysis, they eliminated the sentiments that were vague in representing sentiments. By increasing the sample size, system performance is improved although when the dataset is too large, only increasing the size of training set of data is may not be enough. The eWOM (electronic Word Of Mouth) analysis for specific brands by Jansen, Zhang, Sobel & Chowdhury[4] shows that customer brand perceptions and purchasing decisions appear increasingly influenced by Web communications and social networking services, as consumers increasingly use these

communication technologies for trusted sources of information, insights, and opinions. Apache Hadoop (an open source Java framework for processing and querying vast amounts of data on large clusters of commodity hardware) is used for storing and processing huge amount of data. Hadoop also deals with structured and semi-structured data, XML/JSON files, for example.

1.3 Related Works

A lot of studies have been conducted that can be considered to be related to the discovery of trends in bullying on Twitter. In the paper [1], the authors present the results of machine learning algorithms for classifying the sentiment of Twitter messages. They classify tweets either as positive or negative with respect to specific emoticons found in the Twitter messages.

In paper [4], the authors study how microblogging can be used for sentiment analysis purposes. They show how to use Twitter as a corpus for sentiment analysis and opinion mining. They use a dataset formed of collected messages from Twitter.

Collected tweets were carefully analyzed to identify the most popular hashtags and named entities used within cyberbullying tweets in paper [7].

There has been a paper published on sentiment analysis on bullying and they have identified seven most common emotion expressed in bullying traces; Anger, Embarrassment, Empathy, Fear, Pride, Relief and Sadness by Zhu, Xu and Bellmore [6].

In [24], the authors illustrate a sentiment analysis approach to extract sentiments associated with negative or positive polarity of specific subjects in a document, instead of classifying the whole document as positive or negative. The essential issues in sentiment analysis are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject.

It has been observed that the accuracy of term search extraction of data is not too high and still there have been a lot of study going to improve the pattern or trend analysis as the data spread

over twitter is not always informative enough to predict the trends immediately. In the study conducted in [25] the authors sought to identify the hashtags that are associated with posts about bullying and to determine whether there are commonalities among the different ways in which hashtags are be used. They combined social science and machine learning methods to public mentions of bullying that contain hashtags on Twitter between January 1, 2012 and December 31, 2012. This approach allows for a large-scale, real-time, multi perspective account of how bullying is represented within social media. Their primary goal was to identify the hashtags most frequently associated with Twitter posts that use bullying keywords in 2012. The second goal was to categorize the different types of bullying hashtags that are used by evaluating their different intents. In doing so, we would learn whether hashtags are used differently despite focusing on the shared topic of bullying.

1.4 Motivation

As technology advances, it increasingly allows people to stay connected as one from all over the world, and that too at real time. People have become more vocal over the Internet, sharing their views on social media. Both positive and negative occurrences and viewpoints come to surface. Anyone on the internet can get to know anything that is happening on any part of the world within seconds. A single upload of a picture or a single ‘status’ update has the ability to create immense ripples all over. This has given birth to a ‘internet community’ that is bigger than any physical community there is, and what this implies is even bigger. Where previously only community members could be expected to look out for each other, now we all can be there for one another. This idea is what has helped us to come up with our project.

Social events, occurrences or movements can get ‘viewed’, ‘shared’ and ‘followed’ by millions within hours, spreading from one person to others like wildfire due to the various social media. We have discovered various social events and social movements through Twitter that became popularized by the use of viral hashtags. Hashtagging a topic means that whatever is being said by whosoever can be viewed just by following that hashtag. The more people that use a particular hashtag, the likelier it becomes for that topic to become one of the top trending topics shown in Twitter, thus making it reachable to even more people. This can be particularly helpful when trying

to bring about positive changes, such as, standing up for what is right, supporting others when in need, at least virtually, even when it is not possible physically, or calling out on what is wrong. There have been instances where even people on the verge of committing suicide, posting what they intended to be their good-bye message, have been talked out of it by people from different parts of the world replying with inspiring and encouraging words to go on living. This profound influence that social media allows is the driving force behind our idea. Our intention is to analyze geo-tagged tweets in Twitter that contain hashtags pertaining to different forms of bullying and harassment so that proper action can be taken against such harmful activities that are poisoning our society. Knowing where these forms of bullying and harassment are taking place ensures that the authorities and helping organizations know exactly where the help is needed, and they can set up required institutions to help the victims. This will give victims the confidence that they are not in this alone, and offenders will also be cautious from fearing widespread recognition. Also, awareness campaigns and social movements catering to the people of each particular region would help diminish the mindsets of the corrupt or at least raise enough awareness so that people are alert and supportive of each other in their communities.

1.4 Alternative approach to using STATISTICA

STATISTICA has several alternatives and one of them is SAS. SAS offers products such as data mining, forecasting, text analytics, statistical analysis, among others. However, it is also known for being a heavy and difficult to install, which makes it more suitable for large companies than start-ups. Like most advanced analytics tools, it is difficult for beginners to understand their way around the software.

One other alternative is IBM's SPSS. It is a tool for predictive and statistical analysis that is designed for small and medium size businesses. Its limitations are when it comes to advanced modelling and development of statistical approaches, such as shutting down when the data is big. Some knowledge of language specific to SPSS is also required to work effectively.

STATISTICA itself has its own problems, one of which we faced when we were building trees from a train set. The program would not accept more than 50 rows. However, it is easy to use, gives great cost value and has an accessible graphical user interface that is superior compared to almost all others. In addition, it allows connection with SQL Server and even Hadoop for manipulating large datasets.

Chapter 2

Preliminaries

In this chapter, we define some basic terminology of Twitter data and data mining that we will use throughout the rest of this thesis. Definitions which are not included in this chapter will be introduced as they are needed. We review, in Section 2.1, some definitions of standard twitter and big data terms. In Section 2.2, we discuss about some special definitions like STATISTICA, text mining and SVD, which are important for the ideas and concepts used in the later parts of this thesis.


2.1 Basic Terminologies

We will be providing the definitions of some common terms in this section.

2.1.1 Tweets and Retweets

A “tweet” is a string of characters or simply a message on twitter restricted to 140 characters. A person needs to open up an account in twitter where he/she can send or receive tweets and also be connected with friends. The idea of tweets began when twitter was developed in 2006, its co-founder Jack Dorsey had imagined it to be an SMS-based communications platform. Friends could keep posting status updates known as tweets to keep tab of each other. The 140 character restriction on tweets is because it was originally designed as an SMS mobile phone based platform. Even though twitter kept growing and is still growing the tweets are confined to 140 characters, one can think of it as a creative constraint. In twitter, connecting with friends mean having to follow people and have followers who can view the posts shared on the timeline and send or receive tweets, retweet a message and so on.

People express their thoughts, opinions and emotions through tweets where one can not only write in plain text but also include URLs and pictures. Twitter users can share news and events and retweet to one another’s post. Retweet is a reposting of a tweet, retweets can be found in one’s timeline, profile and other profile pages on Twitter. The retweet feature quickly helps share a tweet

with users' followers and one can retweet his/her own tweet or someone else's tweet. Retweets look like normal Tweets with the author's name and username next to it, but are distinguished by the **Retweet** icon  and the name of the user who Retweeted the Tweet. Some users can block the option of retweeting their posts so that others cannot retweet. One can see who have retweeted their tweets from the notification tab. There is no limit to the number of times a tweet can be retweeted but only top and most recent 100 people's' tweets will be shown on the Home timeline who retweeted public posts. Currently the limit per day for each account is 1000 direct messages sent, 2400 tweets where retweets are counted as tweets. When hitting a limit, twitter sends an error message that limit has been reached to try messaging or sending tweets after the limit period has elapsed.

2.1.2 Hashtags

Anyone using twitter is familiar with the term “hashtag” which is actually used to simply categorize a tweet's topic. Hashtags make it easier for users to identify a message under some specific theme or content. Hashtags are not case-sensitive so if a user searches for a specific tag on the search box for example if the user searches for “#Dhaka”, all the recent tweets or posts related to Dhaka or containing the tag #Dhaka will appear as the results. The result page has different filtering options for the given list of results, the default is Top. One can choose live feed as well as feeds of tweets that include links to current news stories, photos and videos. There are filter options such as from everyone, from people you follow, near you, etc. One of the most amusing things about hashtags is that it allows to create communities of people who are interested in similar topic by making it easy to share related information.

Any user can create and use the hashtag to his/her tweet in support of an event or to show interest in a specific field or to highlight some news. Hashtags recently are also used to promote brands, praise people, express a happy or sad or any kind of emotion and highlight breaking news. Some examples of hashtags related to bullying may include #racism, #stopbullying, #violenceagainstwomen, #xenophobia, #islamophobia and so on. American Express took the advantage of the hashtag trend and collaborated with twitter in February 2013 to allow users to pay for discounted goods online by tweeting a special hashtag. All the American Express cardholders can sync their card with Twitter and pay for offers by tweeting; American Express

tweets a response to the member that confirms the purchase. Hashtags are also used by users on Twitter to review products and criticize about big companies letting others know for example McDonald's created #McDStories to share positive posts but within two hours the restaurant received negative tweets destroying the marketing efforts. A hashtag can easily become popular and get viral over a very short span of time and also the same way can lose its popularity depending on the lifespan of an ongoing event or hype.

2.1.3 Big Data or Data Mining

A very trending term "Big Data" has been heard over the past few years that define large volume of data. Everyday we are storing tons of data in our warehouses and the amount of data is increasing as each passing day so to analyze those we need Big Data Analysis. The importance of Big Data is not the volume of data but what is being done with that data is significant. Big data can be categorized by the following characteristics:

- Volume- The amount of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not
- Velocity- The speed at which data is gathered and processed to meet the demands and challenges that lie in the path of growth and development
- Variability- The inconsistency of data can hamper the processes to handle and manage it
- Veracity- The quality of captured data can vary greatly, affecting the accurate analysis
- Variety- The type and nature of data, this helps people who analyse it to effectively use the resulting insight

Big data analytics can help determine root causes of failures, issues and defects in near-real time, it can help detect any fraudulent activity before it affects one's' organization. It is widely used for business organizations for market research, to analyze customer's preferences and products are modified or made accordingly. The degree of complexity within the datasets is also important,

valuable and complex datasets naturally tend to grow fast and so big that the data becomes massive. It helps employers to take a wiser business decisions depending on a greater predictive analysis such as optimizing operations, preventing threats and frauds, capitalizing on new sources of revenue in large companies.

2.1.4 Tweepy

Tweepy is one of the python libraries which is open-sourced, hosted on GitHub and enables python to communicate with twitter platform and give the access to use its API. Tweepy can access Twitter using newer authentication method, OAUTH. This authentication method has consumer key and access tokens that are provided from the app created at dev.twitter.com for better security. It is possible to get any object with Tweepy that the official twitter API offers.

Streaming twitter data is one of the main usages of Tweepy and the key component is the StreamListener object which monitors the tweets in real time and catches them.

2.2 Technical Terminologies

In this section we will be defining specific terms related to the technical perspective of the project.

2.2.1 STATISTICA

STATISTICA is an analytic software used for data analysis, data mining, statistics, clustering, database management, and various other applications. It was first released by StatSoft in 1991, acquired by Dell in 2014 and now owned by TIBCO Software Inc. It is written in C++ programming language and is renowned for its ease of use among analysts. It offers a concise and user-friendly graphical interface to its extensive array of statistical and data mining tools. STATISTICA is also preferred by analysts because it lets them backtest(build models to test on older cases) and also to modify existing models, which enables them to refine those models. STATISTICA supports Python and R scripts, which are now at the forefront of the Big Data industry.

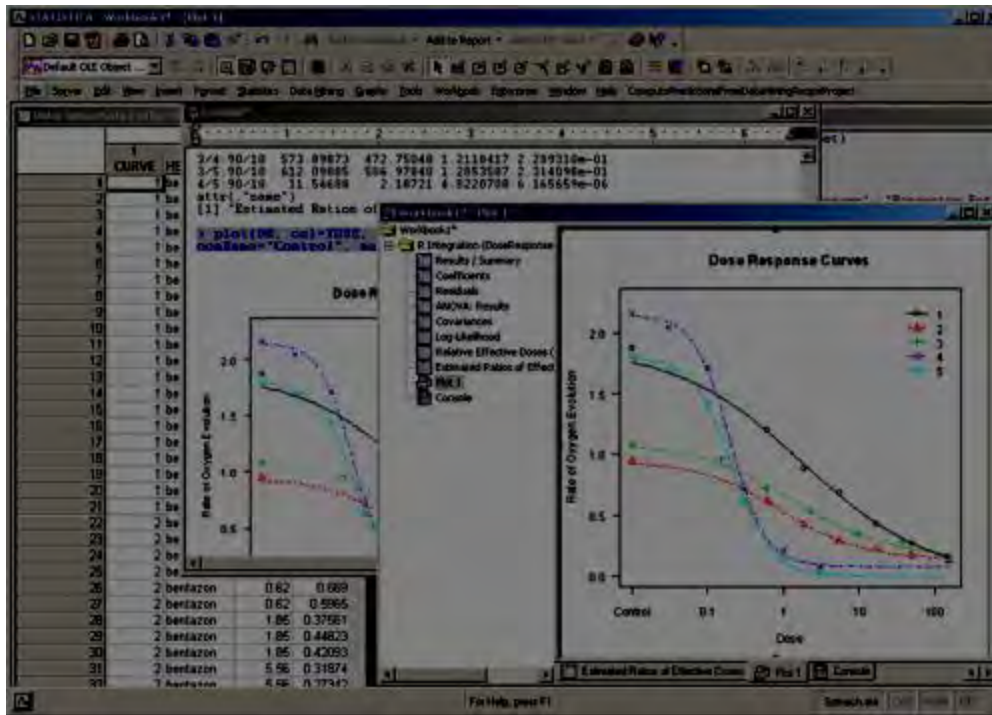


Fig: A look at the STATISTICA graphical interface

Interacting with Statistica is essentially similar to working with any other program designed for Windows. However, there are some differences. Virtually any statistic that we wish to perform can be accomplished in combination with pointing and clicking on the menus and various interactive dialog boxes. Statistica, like many other packages, can be accessed by programming short scripts, instead of pointing and clicking. In this point and click environment one must often navigate through many layers of menu items before encountering the required option. It has a spreadsheet like interface. In Statistica, there are a number of menu options relating to statistics. There are also shortcut icons on the vertical and horizontal toolbar. These serve as quick access to often used options. Holding your mouse over one of these icons for a second or two will produce a short function description for that icon. Data can either be entered manually, or it can be read from an existing data file. The [Open Data] option will launch a dialog box that can be used to open existing data files in a native Statistica format. Like other application packages (e.g., WordPerfect, Excel, etc.) Statistica also has its own format for saving data. In this case, the accepted extension for any file saved using the proprietary format is ".sta". So, one can have a datafile saved as "data1.sta".

The format is not readable with a text editor (e.g., Notepad), it is a binary format. The benefits are that all formatting changes are maintained and the file can be read faster, hence the [Open Data] option. It is specifically meant for files saved in the SPSS format. The second option, [Import Data], as the name suggests is to read files that are not in the statistica format. These include ASCII, or text, datafiles and various spreadsheet formats. By clicking on [Import Data => Quick] one can specify the format of the data file to be read. That is, a new dialog box will appear, and one can scroll through the list of available file types - explore this dialog box. The process is seamless, however, reading ASCII files requires the user to have adequate knowledge about the format of the datafile. Otherwise, one is likely to get stuck in the process of reading. There are a number of acceptable formats - comma separated, space separated, semicolon separated, tab separated, and even a user defined format.

According to the electronic manual that comes with the software, the procedures used in product development generally involve two steps:

1. predicting responses on the dependent, or Y variables, by fitting the observed characteristics of the product using a regression equation based on the levels of the independent, or X variables
2. finding the levels of the X variables that simultaneously produce the most desirable predicted responses on the Y variables.

Many of the statistics are provided in the form of eye-catching infographics that can be used in reports or presentations. Statistics can be exported and displayed in bar chart, line graph or table format – the user can decide which display they prefer, allowing to adapt individual statistics accordingly.

2.2.2 Text Mining

Text Mining enables users to process unstructured data in a way which brings the data up to shape and makes it suitable for the data mining algorithms that STATISTICA provides. It can turn text into numerical values which makes it accessible to features like predictive analytics. Text mining helps determine the relationships between different values or variables in a project.

Unstructured text is very common, and in fact may represent the majority of information available to a particular research or data mining project. Text mining is the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to accuracy, completeness, consistency, uniqueness, and timeliness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).

Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods. A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted.

Traditional keyword search retrieves all the documents that contain the keywords specified.

While that may be useful, the user still has to read all those documents to find out whether they actually contain any information that's relevant to the search. Text mining software is very different, because it reads and analyzes the documents on behalf of the user. It can understand real meanings thanks to sophisticated Natural Language Processing (NLP) algorithms, which allow it to recognize similar concepts – even if they've been expressed in very different ways, or with different spellings. A search using text mining will identify facts, relationships and assertions that would otherwise remain buried in a mass of 'big data'.

The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to

the various data mining (statistical and machine learning) algorithms. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. Hence, one can analyze words, clusters of words used in documents, etc., or one could analyze documents and determine similarities between them or how they are related to other variables of interest in the data mining project. In the most general terms, text mining will "turn text into numbers" (meaningful indices), which can then be incorporated in other analyses such as predictive data mining projects, the application of unsupervised learning methods (clustering), etc.

2.2.3 Typical Applications for Text Mining

Analyzing open-ended survey responses: In survey research (e.g., marketing), it is not uncommon to include various open-ended questions pertaining to the topic under investigation. The idea is to permit respondents to express their "views" or opinions without constraining them to particular dimensions or a particular response format. This may yield insights into customers' views and opinions that might otherwise not be discovered when relying solely on structured questionnaires designed by "experts." For example, you may discover a certain set of words or terms that are commonly used by respondents to describe the pro's and con's of a product or service (under investigation), suggesting common misconceptions or confusion regarding the items in the study.

Automatic processing of messages, emails, etc: Another common application for text mining is to aid in the automatic classification of texts. For example, it is possible to "filter" out automatically most undesirable "junk email" based on certain terms or words that are not likely to appear in legitimate messages, but instead identify undesirable electronic mail. In this manner, such messages can automatically be discarded. Such automatic systems for classifying electronic messages can also be useful in applications where messages need to be routed (automatically) to the most appropriate department or agency; e.g., email messages with complaints or petitions to a municipal authority are automatically routed to the appropriate departments; at the same time, the emails are screened for inappropriate or obscene messages, which are automatically returned to the sender with a request to remove the offending words or content.

Analyzing warranty or insurance claims, diagnostic interviews, etc. In some business domains, the majority of information is collected in open-ended, textual form. For example, warranty claims or initial medical (patient) interviews can be summarized in brief narratives, or when you take your

automobile to a service station for repairs, typically, the attendant will write some notes about the problems that you report and what you believe needs to be fixed. Increasingly, those notes are collected electronically, so those types of narratives are readily available for input into text mining algorithms. This information can then be usefully exploited to, for example, identify common clusters of problems and complaints on certain automobiles, etc. Likewise, in the medical field, open-ended descriptions by patients of their own symptoms might yield useful clues for the actual medical diagnosis.

Investigating competitors by crawling their web sites. Another type of potentially very useful application is to automatically process the contents of Web pages in a particular domain. For example, you could go to a Web page, and begin "crawling" the links you find there to process all Web pages that are referenced. In this manner, you could automatically derive a list of terms and documents available at that site, and hence quickly determine the most important terms and features that are described. It is easy to see how these capabilities could efficiently deliver valuable business intelligence about the activities of competitors.

2.2.4 Approaches to Text Mining

To reiterate, text mining can be summarized as a process of "numericizing" text. At the simplest level, all words found in the input documents will be indexed and counted in order to compute a table of documents and words, i.e., a matrix of frequencies that enumerates the number of times that each word occurs in each document. This basic process can be further refined to exclude certain common words such as "the" and "a" (stop word lists) and to combine different grammatical forms of the same words such as "traveling," "traveled," "travel," etc. This is known as 'stemming'. However, once a table of (unique) words (terms) by documents has been derived, all standard statistical and data mining techniques can be applied to derive dimensions or clusters of words or documents, or to identify "important" words or terms that best predict another outcome variable of interest.

Using well-tested methods and understanding the results of text mining. Once a data matrix has been computed from the input documents and words found in those documents, various well-known analytic techniques can be used for further processing those data including methods for clustering, factoring, or predictive data mining. "Black-box" approaches to text mining and extraction of concepts. There are text mining applications which offer "black-box" methods to

extract "deep meaning" from documents with little human effort (to first read and understand those documents). These text mining applications rely on proprietary algorithms for presumably extracting "concepts" from text, and may even claim to be able to summarize large numbers of text documents automatically, retaining the core and most important meaning of those documents. While there are numerous algorithmic approaches to extracting "meaning from documents," this type of technology is very much still in its infancy, and the aspiration to provide meaningful automated summaries of large numbers of documents may forever remain elusive. We urge skepticism when using such algorithms because 1) if it is not clear to the user how those algorithms work, it cannot possibly be clear how to interpret the results of those algorithms, and 2) the methods used in those programs are not open to scrutiny, for example by the academic community and peer review and, hence, we simply don't know how well they might perform in different domains. For example, one can try the various automated translation services available via the Web that can translate entire paragraphs of text from one language into another. And then translate some text, even simple text, from their native language to some other language and back, and review the results. Almost every time, the attempt to translate even short sentences to other languages and back while retaining the original meaning of the sentence produces humorous rather than accurate results. This illustrates how difficult it is to automatically interpret the meaning of text.

Text mining as document search: There is another type of application that is often described and referred to as "text mining" - the automatic search of large numbers of documents based on keywords or key phrases. This is the domain of, for example, the popular internet search engines that have been developed over the last decade to provide efficient access to Web pages with certain content. While this is obviously an important type of application with many uses in any organization that needs to search very large document repositories based on varying criteria, it is very different from what we have worked with.

2.2.5 Issues and Considerations for "Numericizing" Text

Large numbers of small documents vs. small numbers of large documents: If one's intent is to extract "concepts" from only a few documents that are very large (e.g., two lengthy books), then

statistical analyses are generally less powerful because the "number of cases" (documents) in this case is very small while the "number of variables" (extracted words) is very large.

Excluding certain characters, short words, numbers, etc: Excluding numbers, certain characters, or sequences of characters, or words that are shorter or longer than a certain number of letters can be done before the indexing of the input documents starts. One may also want to exclude "rare words," defined as those that only occur in a small percentage of the processed documents.

Include lists, exclude lists (stop-words): Specific list of words to be indexed can be defined; this is useful when you want to search explicitly for particular words, and classify the input documents based on the frequencies with which those words occur. Also, "stop-words," i.e., terms that are to be excluded from the indexing can be defined. Typically, a default list of English stop words includes "the", "a", "of", "since," etc., i.e., words that are used in the respective language very frequently, but communicate very little unique information about the contents of the document.

Synonyms and phrases. Synonyms, such as "sick" or "ill", or words that are used in particular phrases where they denote unique meaning can be combined for indexing. For example, "Microsoft Windows" might be such a phrase, which is a specific reference to the computer operating system, but has nothing to do with the common use of the term "Windows" as it might, for example, be used in descriptions of home improvement projects.

Stemming algorithms: An important pre-processing step before indexing of input documents begins is the stemming of words. The term "stemming" refers to the reduction of words to their roots so that, for example, different grammatical forms or declinations of verbs are identified and indexed (counted) as the same word. For example, stemming will ensure that both "traveling" and "traveled" will be recognized by the text mining program as the same word.

Support for different languages: Stemming, synonyms, the letters that are permitted in words, etc. are highly language dependent operations. Therefore, support for different languages is important.

2.2.6 Transforming Word Frequencies

Once the input documents have been indexed and the initial word frequencies (by document) computed, a number of additional transformations can be performed to summarize and aggregate the information that was extracted.

Log-frequencies: First, various transformations of the frequency counts can be performed. The raw word or term frequencies generally reflect on how salient or important a word is in each document. Specifically, words that occur with greater frequency in a document are better descriptors of the contents of that document. However, it is not reasonable to assume that the word counts themselves are proportional to their importance as descriptors of the documents. For example, if a word occurs 1 time in document A, but 3 times in document B, then it is not necessarily reasonable to conclude that this word is 3 times as important a descriptor of document B as compared to document A. Thus, a common transformation of the raw word frequency counts (wf) is to compute:

$$f(wf) = 1 + \log(wf), \text{ for } wf > 0$$

This transformation will "dampen" the raw frequencies and how they will affect the results of subsequent computations.

Binary frequencies: Likewise, an even simpler transformation can be used that enumerates whether a term is used in a document; i.e.:

$$f(wf) = 1, \text{ for } wf > 0$$

The resulting documents-by-words matrix will contain only 1s and 0s to indicate the presence or absence of the respective words. Again, this transformation will dampen the effect of the raw frequency counts on subsequent computations and analyses.

Inverse document frequencies: Another issue that one may want to consider more carefully and reflect in the indices used in further analyses are the relative document frequencies (df) of different words. For example, a term such as "guess" may occur frequently in all documents, while another term such as "software" may only occur in a few. The reason is that we might make "guesses" in various contexts, regardless of the specific topic, while "software" is a more semantically focused term that is only likely to occur in documents that deal with computer software. A common and very useful transformation that reflects both the specificity of words (document frequencies) as

well as the overall frequencies of their occurrences (word frequencies) is the so-called inverse document frequency (for the i'th word and j'th document):

$$idf(i, j) = \begin{cases} 0 & \text{if } wf_{i,j} = 0 \\ (1 + \log(wf_{i,j})) \log \frac{N}{df_i} & \text{if } wf_{i,j} \geq 1 \end{cases}$$

In this formula, N is the total number of documents, and df_i is the document frequency for the i'th word (the number of documents that include this word). Hence, it can be seen that this formula includes both the dampening of the simple word frequencies via the log function (described above), and also includes a weighting factor that evaluates to 0 if the word occurs in all documents ($\log(N/N)=0$), and to the maximum value when a word only occurs in a single document ($\log(N/1)=\log(N)$). It can easily be seen how this transformation will create indices that both reflect the relative frequencies of occurrences of words, as well as their semantic specificities over the documents included in the analysis.

2.2.7 Latent Semantic Indexing via Singular Value Decomposition

As described above, the most basic result of the initial indexing of words found in the input documents is a frequency table with simple counts, i.e., the number of times that different words occur in each input document. Usually, we would transform those raw counts to indices that better reflect the (relative) "importance" of words and/or their semantic specificity in the context of the set of input documents.

A common analytic tool for interpreting the "meaning" or "semantic space" described by the words that were extracted, and hence by the documents that were analyzed, is to create a mapping of the word and documents into a common space, computed from the word frequencies or transformed word frequencies (e.g., inverse document frequencies). In general, here is how it works:

Suppose we indexed a collection of customer reviews of their new automobiles (e.g., for different makes and models). One may find that every time a review includes the word "gas-mileage," it also includes the term "economy." Further, when reports include the word "reliability" they also include the term "defects" (e.g., make reference to "no defects"). However, there is no consistent pattern regarding the use of the terms "economy" and "reliability," i.e., some documents include either one or both. In other words, these four words "gas-mileage" and "economy," and "reliability" and "defects," describe two independent dimensions - the first having to do with the overall

operating cost of the vehicle, the other with the quality and workmanship. The idea of latent semantic indexing is to identify such underlying dimensions (of "meaning"), into which the words and documents can be mapped. As a result, we may identify the underlying (latent) themes described or discussed in the input documents, and also identify the documents that mostly deal with economy, reliability, or both. Hence, we want to map the extracted words or terms and input documents into a common latent semantic space.

Singular value decomposition (SVD): The use of singular value decomposition in order to extract a common space for the variables and cases (observations) is used in various statistical techniques, most notably in Correspondence Analysis. The technique is also closely related to Principal Components Analysis and Factor Analysis. In general, the purpose of this technique is to reduce the overall dimensionality of the input matrix (number of input documents by number of extracted words) to a lower-dimensional space, where each consecutive dimension represents the largest degree of variability (between words and documents) possible. Ideally, you might identify the two or three most salient dimensions, accounting for most of the variability (differences) between the words and documents and, hence, identify the latent semantic space that organizes the words and documents in the analysis. In some way, once such dimensions can be identified, you have extracted the underlying "meaning" of what is contained (discussed, described) in the documents.

2.2.8 Incorporating Text Mining Results in Data Mining Projects

After significant (e.g., frequent) words have been extracted from a set of input documents, and/or after singular value decomposition has been applied to extract salient semantic dimensions, typically the next and most important step is to use the extracted information in a data mining project.

Graphics (visual data mining methods). Depending on the purpose of the analyses, in some instances the extraction of semantic dimensions alone can be a useful outcome if it clarifies the underlying structure of what is contained in the input documents. For example, a study of new car owners' comments about their vehicles may uncover the salient dimensions in the minds of those drivers when they think about or consider their automobile (or how they "feel" about it). For marketing research purposes, that in itself can be a useful and significant result. You can use the

graphics (e.g., 2D scatter plots or 3D scatter plots) to help you visualize and identify the semantic space extracted from the input documents.

Clustering and factoring. You can use cluster analysis methods to identify groups of documents (e.g., vehicle owners who described their new cars), to identify groups of similar input texts. This type of analysis also could be extremely useful in the context of market research studies, for example of new car owners. You can also use Factor Analysis and Principal Components and Classification Analysis (to factor analyze words or documents).

Predictive data mining. Another possibility is to use the raw or transformed word counts as predictor variables in predictive data mining projects.

Chapter 3

Design and Implementation

In this chapter we will be describing the full approach and method of implementation for our thesis project. We have provided screenshots of the methods used and described the procedures of how they have been used. Each section describes about specific tasks that were carried out for completing the project.

3.1 Implementation methodology

We chose to work with the high level programming language Python for this study, the first step in which was extracting tweets from Twitter, which consisted of particular hashtags. Hashtags are words or phrases in users' tweets which have the hashtag symbol (#) preceding them; they are used to categorize those tweets so that they can be searched for more easily.

After downloading Python, the first step in the implementation process was to create an application from Twitter's official developer page. We were then provided with a few unique credentials: an access token, a secret access token, a consumer access key and a secret consumer access key, which we later used in the code to extract tweets from Twitter. We endeavored to pick a suitable API for our project by testing a few known ones out, namely 'Twython', 'Tweepy' and Twitter's own Python-based API, 'REST API' specifically. The REST API and Tweepy enabled us to stream tweets according to specific terms, geo locations, etc. hence, we decided to use that.

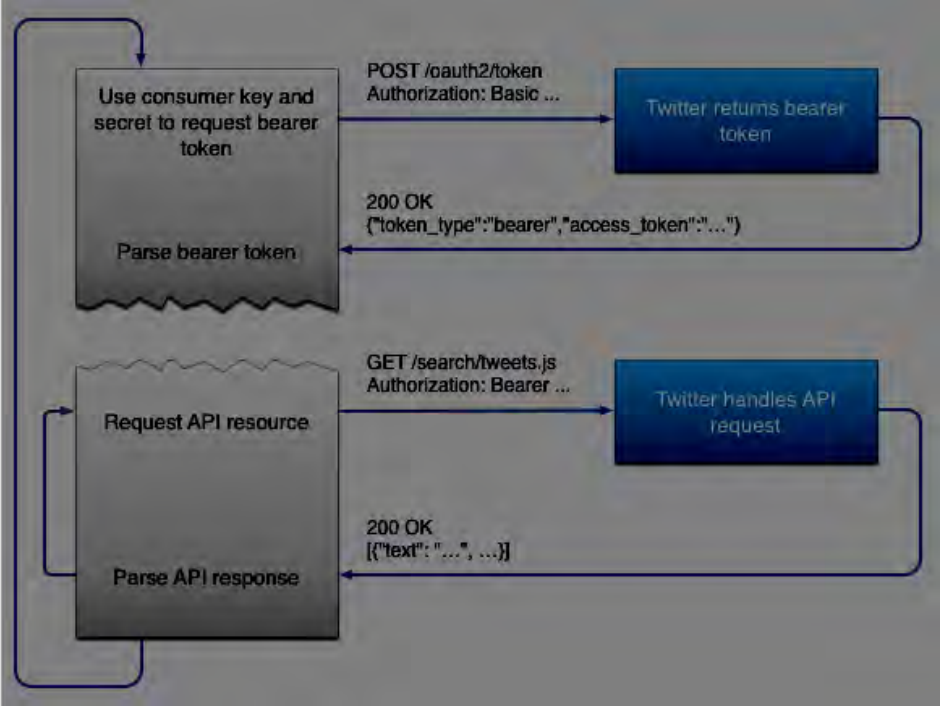


Fig: How Twitter API operates

3.2 Data extraction

The data were derived from the public Twitter streaming API, streaming periodically for a certain number of days and at different times to see the changes in the posts. Using the public streams available through the API (see <https://dev.twitter.com/docs/streaming-apis>) and connecting it to python, tweets were extracted. To meet our aims, relevant features of the data present within the 140 characters in discrete Twitter posts were coded and analyzed.


```
1 import os
2 import json
3
4 from twitter import Api
5
6 ACCESS_TOKEN = "851654330671394817-HEwujqHKMmZ1Ghy1JhNOBt5scOvXYoL"
7 ACCESS_TOKEN_SECRET = "PXFp8I3vnc1fBoXqs5fEC7BV1V1tVsgZ3lvB0TA0crjL6"
8 CONSUMER_KEY = "MpjcaBiZaiNqxqJmNDFONAsnx8"
9 CONSUMER_SECRET = "S8WAgvFHR5y4e4fc3pIePk7QaXKAXdACC5KqqYvMU9zmkLyLRm"
10
11 api = Api(CONSUMER_KEY,
12          CONSUMER_SECRET,
13          ACCESS_TOKEN,
14          ACCESS_TOKEN_SECRET)
15
16 search = api.GetSearch(term='#racism', lang='en', result_type='mixed')
17
18 def main():
19     with open('outputIB58.txt', 'a', encoding='UTF-8') as f:
20         for t in search:
21             print(t.user.screen_name + ' (' + t.created_at + ')')
22             print(t.text.encode('utf-8'))
23             print('')
24
25 if __name__ == '__main__':
26     main()
27
28
29
```

Fig. Twitter REST API Python code for extracting tweets by hashtag

A difficulty we faced while writing the code for data extraction was changes in syntax. Most of the content available online was comprised of terms used in previous versions of the API which were not operational anymore. However, after understanding the documentation properly, we were successfully able to overcome this hurdle and write a working code for extraction. We ran multiple searches on different hashtags (for example, #racism, #sexism, etc.). A constraint we adhered to in the code was that we only extracted tweets which were written in English to increase the reliability of the study. All the date and time information we obtained through the APIs were in Greenwich Mean Time (GMT).

We also could use the latitude, longitude to search for tweets confined between specific latitude and longitude by using Tweepy. A comma-separated list of longitude, latitude pairs specifying a set of bounding boxes to filter Tweets by, only the geolocated tweets falling in the specified pair will be extracted, user's location field is not used to filter tweets. The code that was used to extract

tweets for specified latitude and longitude is given below, Tweepy follows the format [SWLongitude, SWLatitude, NWLongitude, NWLatitude].

```
tweets.py
import tweepy
import csv
auth = tweepy.auth.OAuthHandler('Dif7bkGVnzTLx08JVk5znoDFT', 'kM77Y8Qd1V3wT1i1nH0YwVnsfHpVylXIGi2HFkwZTaz393JX9Y')
auth.set_access_token('2263059900-8znEvtexFnnixy4FCVKPitjywr60mDh07j3n9n6', 'T2ssJ71zinn3w13o22wh5rSLZz4KXVJMUZxbvn42kwE91')
api = tweepy.API(auth)

csvFile = open('ThaiCyberbullying.csv', 'a')

csvWriter = csv.writer(csvFile)

tweet = tweepy.Cursor(api.search,
                       q = "cyberbullying",
                       since = "2017-08-07",
                       until = "2017-08-09",
                       location = "99.777832,13.292743,102.678223,17.276563",
                       lang = "en").items()

csvWriter.writerow([tweet.created_at, tweet.text.encode('utf-8')])
print(tweet.created_at, tweet.text)
csvFile.close()
```

Fig. Tweepy code for extracting tweets by location

We needed to format all the files using Microsoft Excel and WPS Spreadsheets. We needed to separate the screen names from the dates and times, bring each of the screen name, date and time, and tweets to single row and delete two characters which were appearing as a prefix of each tweet (“b”). By using tools like ‘Autofilter’ and ‘Text to Columns’, we fixed all those errors.

Examples of the extracted tweets in csv format using twitter API are as follows:

How the csv filed looked before formatting:

Python 3.6.1 (v3.6.1:69c0db5)									
A	B	C	D	E	F	G	H	I	J
Python 3.6.1 (v3	Mar 21 2017	17:54:52)	[MSC v.1900 32 bit (Intel)]	on win32					
Type "copyright" credits or "license()" for more information.									
>>>									
#ERROR!									
CMCates (Sat Aug 05 14:17:26 +0000 2017)									
b'RT @Lily_Bell82: My doctor just told me I'm the strongest woman she's ever known \xf0\x9f\x92\xaa \n\nSuck it haters! \xf0\x9f\xa4\x97\n#DomesticViolence \n#VictimBlam									
relpolfo (Sat Aug 05 14:08:06 +0000 2017)									
b'RT @EVERALDATALARGE: More than 50 Australian #women die as result of #DomesticViolence annually & we do little. Nutters make a bomb & we sta\xe2\x80\x									
endvawnetwork (Sat Aug 05 14:05:04 +0000 2017)									
b'Donald Trump i #DomesticViolence President #VAW https://t.co/KSQf3NoD3Y via @thedailybanter'									
NorthernOvation (Sat Aug 05 14:04:04 +0000 2017)									
b'One woman's i heroes and love \nhttps://t.co/asplQnbXmS\xe2\x80\xa6 https://t.co/RNJqWP2Sgd"									
WEPHamFul (Sat Aug 05 13:52:46 +0000 2017)									
b'RT @UKSAYSNOMORE: The Other Wounded Warriors: #domesticviolence and the military. Thank you for sharing @McleodHera \nhttps://t.co/vvik54g\xe2\x80\xa6'									
FaithAtheismNub (Sat Aug 05 13:47:03 +0000 2017)									
+ [Menu] domesticviolence.csv									

BaljitRihal (Mon Jul 31 11:10:29 +0000 2017)									
A	B	C	D	E	F	G	H	I	J
BaljitRihal (Mon Jul 31 11:10:29 +0000 2017)									
b'Disgusted by the moronic & #racist comments directed towards #Sikh referee Sukhbir Singh from #Singapore after the\xe2\x80\xa6 htt									
Tony_Tracy (Tue Aug 01 13:41:00 +0000 2017)									
b'Stand against #racism!\n\nWhat *you* can do TODAY to help put an end to #racist police street checks and carding in\xe2\x80\xa6 https://t									
SimonMoyaSmith (Mon Jul 24 16:46:46 +0000 2017)									
b'Where there is Cleveland @Indians\xe2\x80\xa6 https://t.co/tNVyNQq0dq'									
mutts4me_sherri (Tue Aug 01 14:40:03 +0000 2017)									
b'#Racist #Bigoted #Segregation Minnesota College Creates Group to Promote Acceptance...There's Just One Problem https://t.co/4guvV2S									
civ_works (Tue Aug 01 14:39:44 +0000 2017)									
b'GOP Diversity. pink ties blue ties white guys. #Racist #misogyny #xenophobic #KochSuckers... https://t.co/DIDg6WZ0yb'									
mackette52 (Tue Aug 01 14:39:26 +0000 2017)									
b'@glennbeck "FAKE" CRYIN needs EMPATHY \n\nforgot calling @realDonaldTrump supporters #nazis #brownshirt #racist DARED2question									
FoamingPenguin (Tue Aug 01 14:39:23 +0000 2017)									
b'GOP Diversity. pink ties blue ties white guys. #Racist #misogyny #xenophobic #KochSuckers https://t.co/OcVeqvl1N8"									
+ [Menu] outputRacist.csv									

After formatting the data for moving to hive table easily, the look of the csv file:

Info	Datetime	Tweets	D	E	F	G	H	I	J
ABAonline	2017-08-05 13:2	If your being #bullied online don't keep it to yourself. Here are 5 tips #cyberbullying #bullying https://t.co/XYWb4WtaL"							
BuildingMadison	2017-08-05 15:4	Good for them! 50 #bikers escort #bullied kid to school @CNN https://t.co/FHEzoTN7l #stopbullying #standup #saysomething'							
jakeharris317	2017-08-05 14:3	I challenge all of you out there who get #bullied on a daily basis to stand up and be proud of who you are!xe2x80xa6 https://t.co/eQzJ1tNlwd'							
WedgiesGalore	2017-08-05 14:1	RT @wedgievictim: Girls basketball team seniors hazing a freshman with a thong wedgie \xf0\x9f\x98\xac\xf0\x9f\x98\x82 \n\n#wedgie #hazing #th							
FionaSalsa	2017-08-05 13:5	RT @ABAonline: If your being #bullied online don't keep it to yourself. Here are 5 tips #cyberbullying #bullying https://t.co/XYWb4WtaL"							
ConsultantBob	2017-08-05 13:4	RT @ABAonline: If your being #bullied online don't keep it to yourself. Here are 5 tips #cyberbullying #bullying https://t.co/XYWb4WtaL"							
RebeccaNunes6	2017-08-05 13:3	RT @ABAonline: If your being #bullied online don't keep it to yourself. Here are 5 tips #cyberbullying #bullying https://t.co/XYWb4WtaL"							
TamaraBuffard	2017-08-05 13:2	50 #bikers #escort #bullied #boy to first day of middle #school \nhttps://t.co/so2vdOKdmB'							
laifscanada	2017-08-05 13:2	RT @ABAonline: If your being #bullied online don't keep it to yourself. Here are 5 tips #cyberbullying #bullying https://t.co/XYWb4WtaL"							
ABAonline	2017-08-05 13:2	If your being #bullied online don't keep it to yourself. Here are 5 tips #cyberbullying #bullying https://t.co/XYWb4WtaL"							
dinastavola	2017-08-05 12:2	What a one sided advert! Um what about when ANYONE is being #HARASSED & #BULLIED ? https://t.co/bkKojb9gFP"							
CheawnK	2017-08-05 12:0	RT @Renate12081: 10 year old starts clothing line after being #bullied \xf0\x9f\x91\x8f\xf0\x9f\x8f\xbd Beautiful Young lady #FlexInnHerComplexi							
InAnyEvent_LDN	2017-08-05 9:02	You are not the things your own mind has #bullied you into believing https://t.co/LQof7qyKde'							
letsetsepj	2017-08-05 8:24	RT @wedgievictim: Atomic wedgie for the loser girl \xf0\x9f\x98\x88\xf0\x9f\x98\x82 \n\n#wedgie #panties #bullied							
theidk_	2017-08-05 8:17	RT @wedgievictim: When you lose a bet with your friend and have to get wedgied on camera \xf0\x9f\x99\x83 #wedgie #panties #bullied https://t.c							
theidk_	2017-08-05 8:17	RT @wedgievictim: Surprise thong wedgie \xf0\x9f\x98\xac\xf0\x9f\x98\x82 #wedgie #thong #panties #bullied https://t.co/ScDuAYbdXJ'							
theidk_	2017-08-05 8:17	RT @wedgievictim: Ripping thong wedgie in CVS \xf0\x9f\x98\x82\xf0\x9f\x98\x82\xf0\x9f\x98\x82\n\n#wedgie #thong #panties #bullied https://t.co							
theidk_	2017-08-05 8:17	RT @wedgievictim: Giving ill sis a giant wedgie \xf0\x9f\x98\x82 #wedgie #bullied https://t.co/LZWfvyQdzj'							
theidk_	2017-08-05 8:10	RT @wedgievictim: Atomic wedgie for the loser girl \xf0\x9f\x98\x88\xf0\x9f\x98\x82 \n\n#wedgie #panties #bullied							

Name	Datetime	Tweets	D	E	F	G	H	I
NatashaLeeMay	2017-08-07 16:4	@10kirk25 @UN Women #Sexism & #Misogyny is live & well clearly. Hope you learn the error of your ways &						
ShineJob	2017-08-07 16:4	Sex your baby's Sex a secret. (Gender is how you feel as an adult.) #Facepalm #REKT #Segregation #DivideAndConquerxe2x80						
HerAgenda	2017-08-07 16:3	Can you believe this? A Google engineer used this as an excuse as to why there aren't more women in tech.xe2x80xa6 https://t.c						
womanInTransit	2017-08-07 16:3	#Gender gap is natural #Google employee says in 10-page xe2x80x98internally viralxe2x80x99 memo https://t.co/CAJuYot3W						
npquarterly	2017-08-07 16:3	Social media leaks #Google memo against the companyxe2x80x99s measures to improve #diversity https://t.co/kDIXRvcCWH #s						
Spyparent	2017-08-07 16:2	Listen to @UnSlutProject discuss why boys engage in #sexualbullying https://t.co/8174gAaaNT #harassment #sexism https://t.co/						
RNRVirginia	2017-08-07 16:1	RT @RNRTennessee: Can't make this up!\nBehold the power of cheese - according to #PETA\nCheese = #Sexism?!\nhttps://t.co/						
Mati_Monnot	2017-08-07 16:1	Needs citations. https://t.co/XtjndXITZ6 #GoogleManifesto #sexism #genderequality #googlememo #engineerstryingtobesocialscier						
cavemander17	2017-08-07 16:1	Waiting in line at the Morgan Co AL DMV. Only 2 of 71 inductees into Sports Hall of Fame are women. #sexismxe2x80xa6 https://t.c						
RNRMaryland	2017-08-07 16:1	RT @RNRTennessee: Can't make this up!\nBehold the power of cheese - according to #PETA\nCheese = #Sexism?!\nhttps://t.co/						
BusInclusivity	2017-08-07 16:0	Want to hear your stories about #sexism in \n\npublishing. DM or use hashtag #Itsreal https://t.co/dnMaxg8QxL #sexism #publishin						
AllisonQuient	2017-08-07 16:0	Turning the Other Cheek does not mean enabling evil or ignoring it. #hemeneutics #theology #interpretation #violence #abuse #se						
nfreear	2017-08-07 16:0	RT @ShermaKhambatta: xe2x80x98This book is ammunitionxe2x80x99 writer @AngelaDSaini who exposed #sexism and myth						
RexHeraclius	2017-08-07 16:0	RT @Vlaamselr: This female engineer is totally with you! Stay strong courageous manifesto writer \xf0\x9f\x91\x8d#diversity #goog						
ShineJob	2017-08-07 15:3	it's as if the parents treat penis & vagina different #pinkbluedivide #sexism #rapeculture #paedo https://t.co/bVchHaCAWx"						
JoeTomillionII	2017-08-07 15:1	@StacyOnTheRight The lefts only response when not knowing what their talking about is to shout #racismxe2x80xa6 https://t.co/						
MargevonMarge	2017-08-07 15:0	@BootsUK: apologising for your #sexism isn'txe2x80x99t enough! Commit to a lower cost for morning after pill! #JustSayNon http						
re_learning	2017-08-07 15:0	@BootsUK: apologising for your #sexism isn'txe2x80x99t enough! Commit to a lower cost for the morning after pill! #JustSayNon						
jesswade	2017-08-07 15:0	RT @ShermaKhambatta: xe2x80x98This book is ammunitionxe2x80x99 writer @AngelaDSaini who exposed #sexism and myth						

3.3 STATISTICA Installations

We downloaded STATISTICA 12.5, which we then had to extract to install in our computers. In order to install the software, there was a requirement to provide personal details such as name, email address and country for their records.

3.4 Importing data in STATISTICA

In the next step, the data has been imported in STATISTICA. STATISTICA properly interprets formatted cells (such as 4/17/1999 or \$10) and text values, including extensive in-cell formatting (e.g., RVS tower 120.3MHz). Data files from a wide variety of Windows and non-Windows applications can also be accessed and translated into the STATISTICA format (.sta) using the file import facilities. A wide variety of files are available in the Files of type box. Along with the numerous types of STATISTICA documents, Excel, dBASE, SPSS Portable, Lotus/Quatro Worksheets, Text [formatted and free format text (ASCII)], HTML, and Rich Text Files are available. This ability to specify the exact way in which a file is to be imported is a distinct advantage of using the file import facilities instead of the Clipboard. In addition, the user can access types of data that are not (or not easily) accessible to Clipboard operations. In addition to the file import facilities described above, STATISTICA provides access to virtually all databases (including many large system databases such as Oracle, Sybase, etc.) via STATISTICA Query. Additional import options can be accessed on the Import tab of the Options dialog box. You can specify the manner in which Excel, Text, and HTML files will be imported as well as the maximum rows of data that is retrieved by STATISTICA Query. Note that STATISTICA Query is capable of retrieving data larger than the value you specify here. Once you have reached the maximum row, you will be prompted to continue or stop retrieving data.

Accessing data files larger than the local storage. Note that enterprise versions of STATISTICA offer options to query and access large remote data files in-place (i.e., without having to import the data and create a local copy).

3.5 Frequency

There are various statistical summaries that can be computed for each word (within each document). These are mostly simple transformations of the original word frequencies, in order to achieve more meaningful indices with values and distributions (e.g., of the words across the documents) that are more suitable for subsequent analyses using other statistical or data mining techniques.

Use the options in this group box to choose one of these common transformations (or to use raw word frequencies). When you request the Frequency matrix, or perform singular value decomposition (via the Concept extraction tab), the respective computations and summaries are computed and reported for the chosen transformation only (e.g., singular value decomposition can be performed for the raw Frequency counts, Inverse document frequency statistics, and so on).

Inverse document frequency: This option is selected to analyze and report inverse document frequencies. One issue to consider are the relative document frequencies (df) of different words. For example, a term such as "guess" may occur frequently in all documents, while another term such as "software" may only occur in a few. The reason is that one might make "guesses" in various contexts, regardless of the specific topic, while "software" is a more semantically focused term

that is only likely to occur in documents that deal with computer software. A common and very useful transformation that reflects both the specificity of words (document frequencies) as well as the overall frequency of their occurrences (word frequencies) is the so-called inverse document frequency (for the i'th word and j'th document):

$$idf(i, j) = \begin{cases} 0 & \text{if } wf_{i,j} = 0 \\ (1 + \log(wf_{i,j})) \log \frac{N}{df_i} & \text{if } wf_{i,j} \geq 1 \end{cases}$$

In this formula (see also formula 15.5 in Manning and Schütze, 2002), N is the total number of documents, and df_i is the document frequency for the i'th word (the number of documents that include this word). Hence, it can be seen that this formula includes both the dampening of the simple word frequencies via the log function, and also includes a weighting factor that evaluates to 0 if the word occurs in all documents ($\log(N/N=1)=0$), and to the maximum value when a word only occurs in a single document ($\log(N/1)=\log(N)$). It can easily be seen how this transformation will create indices that both reflect the relative frequencies-of-occurrences of words, as well as their semantic specificities over the documents included in the analysis.

Raw: This is the default selection that enables the user to operate on raw word frequencies collected in the term-document index.

Binary: This option is to analyze and report binary indicators instead of word frequencies. Specifically, this option will simply enumerate whether a term is used in a document; i.e.:

$$f(wf) = 1, \text{ for } wf > 0$$

Where wf stands for word frequency within each document. The resulting documents-by-words matrix will contain only 1s and 0s, to indicate the presence or absence of the respective word. As the other transformations of simple word frequencies, this transformation will dampen the effect of the raw frequency counts on subsequent computations and analyses.

Logarithmic. Select this option button to analyze and report logs of the raw word frequencies. A common transformation of the raw word frequency counts (wf) is to compute:

$$f(wf) = 1 + \log(wf), \text{ for } wf > 0$$

This transformation will dampen the raw frequencies and how they will affect the results of subsequent computations.

List of selected words. This list displays the words that were extracted from the documents and their frequencies (the overall word frequencies as well as document frequencies, i.e., number of documents in which they were found). You can sort by each column in the list of extracted and selected words by clicking on the respective column header. For example, to sort by the word itself, click on the Stem/Phrase column header. Click on the Count header to sort by the total word frequencies (click once to sort in ascending order, click again to sort in descending order).

The Stem/Phrase column lists the terms as they were indexed (stored in the internal database, see also the Introductory Overview), i.e., after stemming. This column will also list phrases (user-defined word combinations that should be treated as a whole), if present.

The entries in the Example column show the shortest original words that were reduced to the respective stem, unless such a word is the stem itself, in which case the entry is empty.

The list’s check box controls near each term enable you to select/deselect some of the words in the index. It is important to distinguish between selected and unselected words vs. indexed and non-indexed words. Words or terms can be indexed in the (internal) database but not selected into the word list from which final results are computed (e.g., singular value decomposition). If the Keep unselected words in database for browsing option on the Advanced tab of the Text Mining dialog box is selected, the list will display all words contained in the term-document index, even the ones that did not pass automatic selection conditions; in this case, you can perform word selection manually. The Count column displays the total word frequencies.

The Files column displays the document frequencies of listed words.

3.6 Obtaining Graphical Representation of Tweets and Retweets

The tweets that were extracted according to specific hashtags from the twitter API were then modified. Each file was organized to hold the three columns of name, datetime and tweets, and divided according to dates into separate files and imported into STATISTICA. These files, which were now of specific hashtags tweeted about on specific dates, were plotted in a histogram, with the y-axis being the time of day and the x-axis being the frequency of the tweets or the number of observations during fixed periods of time. We chose to use histograms to represent our records as we only had one variable to work with: ‘Datetime’. The graphs that were generated were saved in png format. Then, by aggregating all the files with data on a single hashtag but different dates, we produced graphs illustrating the trends of those hashtags for consecutive days. The data was sorted terms of dates, in ascending order.

	Name	Datetime	Tweets
1	RayJorden	13-08-2017	@JodieMarsh would love to work with on anti bullying subjects, we should talk x @BulliesOut #St
2	MadlibBot	13-08-2017	Let’s woul remakes TODAY! #antibullying #stopbullying @stompoutbullying https://t.co/7mz97WVCQE
3	W_Angels_Wings	13-08-2017	RT @MaryLSchmidt: Save kids from #cyber bullies! https://t.co/YA0qBgtsA3 #RRRC #SCBWI #PDF1 #ki
4	javonterose	13-08-2017	All NEW & Revamped Tee’s COMINN SOONN \xf0\x9f\xa5\x9f\xa5\x9f\xa5 #tea
5	rchenerson3	13-08-2017	RT @MMSbluedevels1: 8th grade carnival games! #BT1 #antibullying #teambuilding https://t.co/tmnl
6	Sparkles_Blog	13-08-2017	An Open Letter To My High School Bully... https://t.co/v46sEeHnnV #pbloggers #mentalhealth #antil
7	OCFlowPromo	13-08-2017	RT @LisadaErat: #antibullying film by @NoelGugliemi Now available on @amazon & in your scho
8	AmeliaMaddnes	13-08-2017	I think the whole #antibullying stuff is going a bit far. We need to change what is classified i
9	AngieCakin	13-08-2017	Bullies leave plank of wood nailed to 9-year-old boy\xe2\x80\x99s head https://t.co/7t0sdtwZBk
0	Shellyjo624	13-08-2017	How is your #AntiBullying going? https://t.co/dLHckYcS2
1	aj_jk4	13-08-2017	RT @nickvujicic: In Jesus, we have everything we need to STAND STRONG!\nLet’s end #bullying tog
2	IAMDanBot	13-08-2017	DONT MAKE RACOONS FEEL CONSCIOUS ABOUT THEIR WEIGHT #AntiBullying’
3	DemiNewell	13-08-2017	Words Can Be Very Powerful\nhttps://t.co/61swr6RcOY\n\n#etsyfinds #creativebizho
4	firsttutors	13-08-2017	Interesting aspect of this app is the filter that will use AI to filter out mean and bullying m
5	AnitaNaik	13-08-2017	RT @firsttutors: Interesting aspect of this app is the filter that will use AI to filter out me
6	Roaringpurr	13-08-2017	RT @MaryLSchmidt: Save kids from #cyber bullies! https://t.co/YA0qBgtsA3 \n#RRRC #RPEP #t4us #k
7	MargieKay5	13-08-2017	RT @TimeHathCome: #antibullying Bullied Kid Gets Escort to School From 50 Bikers https://t.co/G:
8	ErolMcbride	13-08-2017	#AntiBullying song https://t.co/2qP6uVxelS’
9	sanshooter_sam	13-08-2017	#antibullying #stopbullying just went out for the first time wearing makeup! It felt awesome! h
0	bullyinguk	13-08-2017	Our #antibullying resources can be used all year round! Plenty to choose from, free to download
1	ThreeCirclesLAC	13-08-2017	RT @bullyinguk: Our #antibullying resources can be used all year round! Plenty to choose from,
2	haandleonit	13-08-2017	bullyinguk: Our #antibullying resources can be used all year round! Plenty to choose from, free

Fig: Workbook containing tweets with hashtag ‘#antibullying’ on 13-08-2017

35	LittleMissFlint	13-08-2017	My mentor and (role) model. One of my biggest cheerleaders...Miss Tracey \n#InWithMari \n#Team/
36	kadikadey	13-08-2017	#heatherheyer\n#politicians # peace #sayhername #charlottesville #charlottesvillerrally #cville :
37	CILIPSLG	13-08-2017	RT @uksla_london: 20 powerful books that tackle the subject of bullying. \n#AntiBullying #Power(
38	judehaste_write	13-08-2017	RT @bullyinguk: Our #antibullying #wristbands are a great addition to any #workplace (minimum of
39	FHS_Liby	13-08-2017	RT @uksla_london: 20 powerful books that tackle the subject of bullying. \n#AntiBullying #Power(
40	fabled_films	13-08-2017	Chk out our piece on @ABC2NEWS discussing #children #reading and #learning! #booklove #sunday #:
41	NocturnalsWorld	13-08-2017	Chk out our piece on @ABC2NEWS discussing #children #reading and #learning! #booklove #sunday #:
42	EHFoundation237	13-08-2017	RT @fabled_films: Chk out our piece on @ABC2NEWS discussing #children #reading and #learning! #:
43	EHFoundation237	13-08-2017	RT @NocturnalsWorld: Chk out our piece on @ABC2NEWS discussing #children #reading and #learning
44	kidsanddreams	13-08-2017	VOTE NOW! Kids & Dreams needs your help. Which "Drean Big" design-A or B-do u like best for
45	OCFlowApparel	13-08-2017	RT @LisadaBrat: #antibullying film by @NoelGugliemi Now available on @amazon & in your scho
46	league_network	14-08-2017	Protecting Your Child from Coaches Who Bully - League Network https://t.co/Bi7eokYpqs #YouthSpo
47	ChrisFore3	14-08-2017	RT @league_network: Protecting Your Child from Coaches Who Bully - League Network https://t.co/Bi7eokYpqs
48	johntrprather	14-08-2017	How to become bully proof https://t.co/Xvg4093cjo #bullying #antibullying #blog #amwriting'
49	Sparkles_Blog	14-08-2017	An Open Letter To My High School Bully... https://t.co/v46sEeHmV #bloggers #mentalhealth #antib
50	SblahSblah8	14-08-2017	A4: #Upstanding is now taught as part of #Antibullying programs. \n\x0\x9f\x8e\xa5 https://t.co/yHtH
51	HeliosHR	14-08-2017	Bullying and Harassment in the workplace. How different are they? Learn more: https://t.co/yHtH
52	javonterose	14-08-2017	All NEW & Revamped Tee's COMINN SOONN \xf0\x9f\x94\xa5\x0\x9f\x94\xa5\x0\x9f\x94\xa5 #tear
53	LifecoachNV16	14-08-2017	I'm working on #antibullying guide for kids & I can use any #donations. Click to Donate: h
54	sns_library	14-08-2017	RT @uksla_london: 20 powerful books that tackle the subject of bullying. \n#AntiBullying #Power(
55	SchoolProgress	14-08-2017	We are currently building up our lists on Twitter! We're all about educating young people on #

Fig: Workbook containing tweets from few of the chosen dates

In the ‘Tweets’ columns of the original files extracted, retweets were the values starting with the phrase “RT @”. We determined the count of the tweets starting with “RT @” and all the other tweets in each date for each hashtag. Then the data was categorized into ‘retweets’ and ‘other tweets’ to determine how many of the data containing those hashtags were original and to ascertain which hashtags had most and least retweets. We divided the data into three different columns: ‘Date’, ‘Retweets’ and ‘Other Tweets’.

	Date	Retweets	Other Tweets
1	13/08/2017	21	24
2	14/08/2017	35	35
3	15/08/2017	57	41
4	16/08/2017	29	33
5	17/08/2017	32	34

Fig: Workbook classifying different kinds of tweets

Chapter 4

Results and Visual Representations

4.1 Extracted Tweets

4.1.1 Tweets with different hashtags related to bullying

Using the twitter API and search filtering we filtered specific terms or hashtags that are related to bullying. We mainly used the tags #cyberbullying, #racism, #sexism, #stopbullying, #antibullying, #harrasment, #bullied, #xenophobia. We used the hashtag #cyberbullying to analyze five countries, like to see the frequency of cyberbullying in countries such as Brazil, France, Thailand, India, Canada and USA using their latitude and longitude.

4.1.2 Tweets containing location or time zone

We have used tweepy to stream and save the tweets that had geotagging enabled, not each and every extracted tweet had the feature enabled. There were quite a lot tweets where the user's geotagging feature was enabled and we could obtain the data. We used two ways to obtain the location data, firstly we used the latitude and longitude parameters during the api.search and then we used the status.user.time_zone method to extract time zones. We used the latitude and longitude of some of the countries from the map and used it for filtering the search terms.

The following screenshots show the example of tweets extracted with time zones for specific term searches:

Datetime	Tweets	ID	Time Zone
2017-08-08 23:5	RT @northforkmary: Case Study in Tragedy https://t.co/l83MidFD4F via @AbuseStoppers\n#childabuse	neenna68	
2017-08-08 23:5	@AlanDersh @GatestoneInst What you have said about #FGM is still #childabuse! #EndFGM #stopbullying	i_o_CPP	Eastern Time (US & Canada)
2017-08-08 23:4	Please watch this #childabuse #childprisoners at the hands of brutal and #apartheid #Israel https://sloughpsc		
2017-08-08 23:4	RT @arbetarbroder: Britain's Hidden #ChildAbuse - London's Orthodox Jewish Community - #Pedo	torrexx2003	Eastern Time (US & Canada)
2017-08-08 23:3	RT @carriamahoney: To live is to suffer, to survive is to find some meaning in the suffering. #FriedrichNietzsche	Cam53535998	
2017-08-08 23:3	RT @peepartist: ELLE Magazine promoting 8 year old drag queen. #childabuse https://t.co/hvX7qbDANotLikeYou		Pacific Time (US & Canada)
2017-08-08 23:3	To live is to suffer, to survive is to find some meaning in the suffering. #FriedrichNietzsche #survivor	carriamahoney	Mountain Time (US & Canada)
2017-08-08 23:2	ELLE Magazine promoting 8 year old drag queen. #childabuse https://t.co/hvX7qbWHbM	peepartist	Atlantic Time (Canada)
2017-08-08 23:2	@StefMacWilliams I remember when this was called #childabuse'	Psyllius	Atlantic Time (Canada)
2017-08-08 23:2	RT @carriamahoney: When a man #beats his boy, he wants a son who won't buck him. He's trying to make a #coward...	Cam53535998	
2017-08-08 23:2	Keeping silent of other's evil doings only ends up poisoning you. #youroutervoice #quote #childabus	carriamahoney	Mountain Time (US & Canada)
2017-08-08 23:0	When a man #beats his boy, he wants a son who won't buck him. He's trying to make a #coward...	carriamahoney	Mountain Time (US & Canada)
2017-08-08 22:5	#ChildAbuse is never OK\nCheck out @berrystreet tips for learning the signs of #abuse\nAll #childr	WSMLLEN	
2017-08-08 22:5	RT @AbusedKids: Light shines the brightest in the dark https://t.co/p55n02BWAAd "Hold On" by @Jc quinnCeNation		Mountain Time (US & Canada)
2017-08-08 22:5	You only live once, make it count! Help #AdultSurvivors of #ChildAbuse via #RemovingChains https://t.co/AbusedKids	AbusedKids	America/New_York
2017-08-08 22:4	#pizzagate #pedogate #jimmysavile #childabuse #abuse\nlink >\n https://t.co/AUhV40aobf https://t.co/9Chriz		Pacific Time (US & Canada)
2017-08-08 22:4	RT @AbusedKids: Light shines the brightest in the dark https://t.co/p55n02BWAAd "Hold On" by @Jc JordanAdams		London
2017-08-08 22:4	RT @AbusedKids: #Runners will you Race Unchained this year to support #ChildAbuse Survivors? MeenaGounder		
2017-08-08 22:4	RT @AbusedKids: Did you do something today to help you reach your life goals? Help #ChildAbuse MeenaGounder		

Fig: Tweets extracted without location filter, displaying the time zone

A	B
Datetime	Tweets
2017-08-08 19:28:28	@FLOTUS @POTUS @SecPriceMD I hope you are more successful with this than with you were with your anti-cyberbullying campaign'
2017-08-08 19:28:01	@FLOTUS @POTUS @SecPriceMD Hows that cyberbullying campaign going? You said you would devout your time to it. Not wlx2lx80'xa6 https://t.co/8BLmp454E'
2017-08-08 18:55:37	@FLOTUS @POTUS @SecPriceMD I hope this is as successful as your anti-cyberbullying campaign! Oh wait...'
2017-08-08 18:26:44	@FLOTUS @POTUS @SecPriceMD if it goes as well as your 'Stop Cyberbullying' campaign, it's gonna suck."
2017-08-08 18:16:34	@FLOTUS @POTUS @SecPriceMD Hope you can accomplish as much as with your powerful anti-cyberbullying campaign. ...oh, lx2lx80'xa6 https://t.co/mnaPc5NUYI'
2017-08-08 17:42:22	@FLOTUS @POTUS @SecPriceMD Yes, thank you for calling attention to this. Another problem: cyberbullying. Please spelx2lx80'xa6 https://t.co/ovXnYgQ0HO'
2017-08-08 17:27:09	@FLOTUS @POTUS @SecPriceMD What happen to your cyberbullying initiative! Your husband need them with the quickness!'
2017-08-08 17:10:23	@FLOTUS What is the status of your #cyberbullying initiative? Here are some great tips from @WebMD: https://t.co/2FMjRCFA2'
2017-08-08 16:31:27	@FLOTUS @POTUS @SecPriceMD Your husband @realDonaldTrump is MAKING ME TAKE DRUGS @FLOTUS - stop him fromlx2lx80'xa6 https://t.co/J3FyaJlFrP'
2017-08-08 16:03:19	@FLOTUS @POTUS @SecPriceMD CYBERBULLYING STARTS AT HOME! The Dump you are having redecorated !!!'
2017-08-08 15:24:55	@FLOTUS @POTUS @SecPriceMD What about cyberbullying? That FatBaby does it daily! Oh, new project? First one not finished... wow MAGA'
2017-08-08 15:21:57	@FLOTUS @POTUS @SecPriceMD Wondering how your #CyberBullying campaign is coming along?'
2017-08-08 15:19:39	@FLOTUS @POTUS @SecPriceMD You seem to choose causes that the prez is struggling with like cyberbullying. Does thelx2lx80'xa6 https://t.co/X1f18c8EU'
2017-08-08 14:23:10	@FLOTUS @POTUS @SecPriceMD What's going on with the Cyberbullying Project? Is your husband the poster child?'
2017-08-08 14:19:03	@FLOTUS #OpioidCrisis let's hope she does a better job on this then she has on #CyberBullying @ChrisChristie @POTUS @SenBookerOffice @CNN'
2017-08-08 13:51:54	@FLOTUS @POTUS @SecPriceMD What about cyberbullying? POTUS is your worst offender. There's no controlling him so itlx2lx80'xa6 https://t.co/5abmpDsDWo'
2017-08-08 13:49:54	@FLOTUS @POTUS @SecPriceMD Your cyberbullying initiative was typical of you -nothing was accomplished. #ImpeachTrump #TrumpCrimeFamily'
2017-08-08 13:34:38	@FLOTUS @POTUS @SecPriceMD Hopefully your contribution today will be more successful than your anti cyberbullying efforts, eh?'
2017-08-08 13:33:52	@FLOTUS @POTUS @SecPriceMD What happened to Cyberbullying?'

Fig: Tweets filtered by specified latitude/longitude: America's coordinates used.

4.2 Frequency of different hashtags

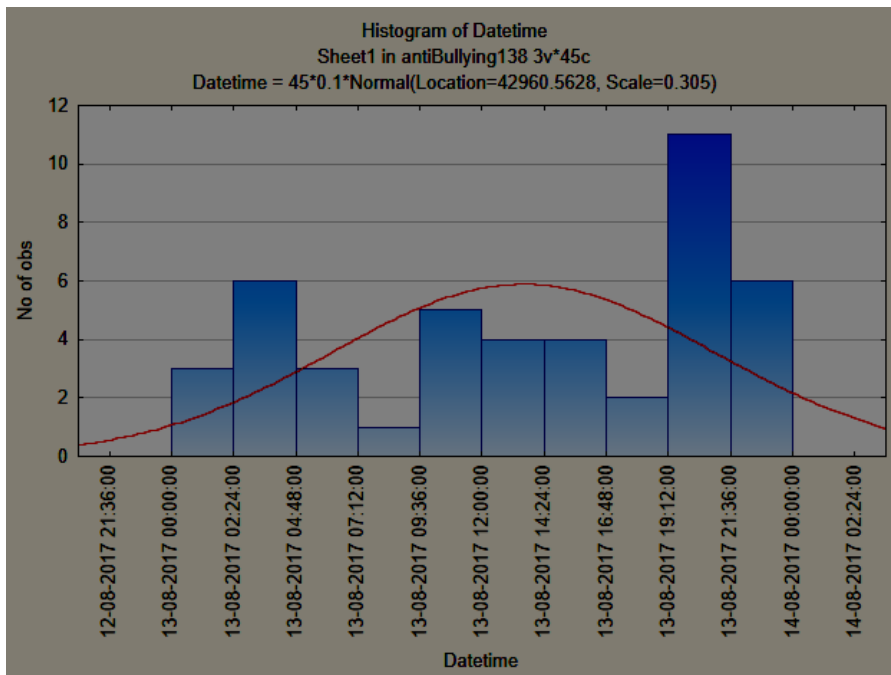


Fig :Frequency of tweets containing #antibullying hashtag on 13-08-2017

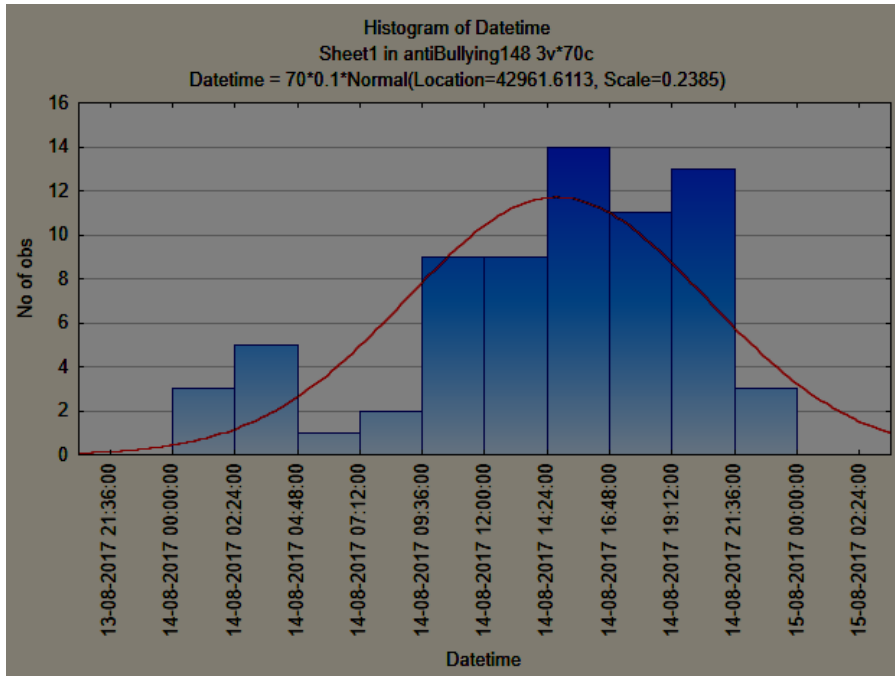


Fig: Frequency of tweets containing #antibullying hashtag on 14-08-2017

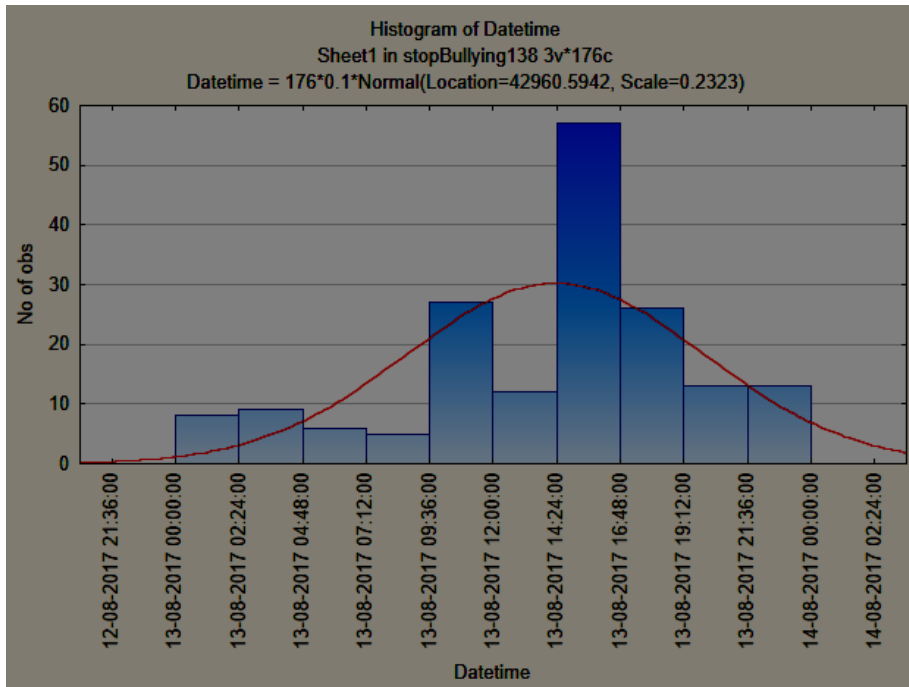


Fig: Frequency of tweets containing #stopbullying hashtag on 13-08-2017

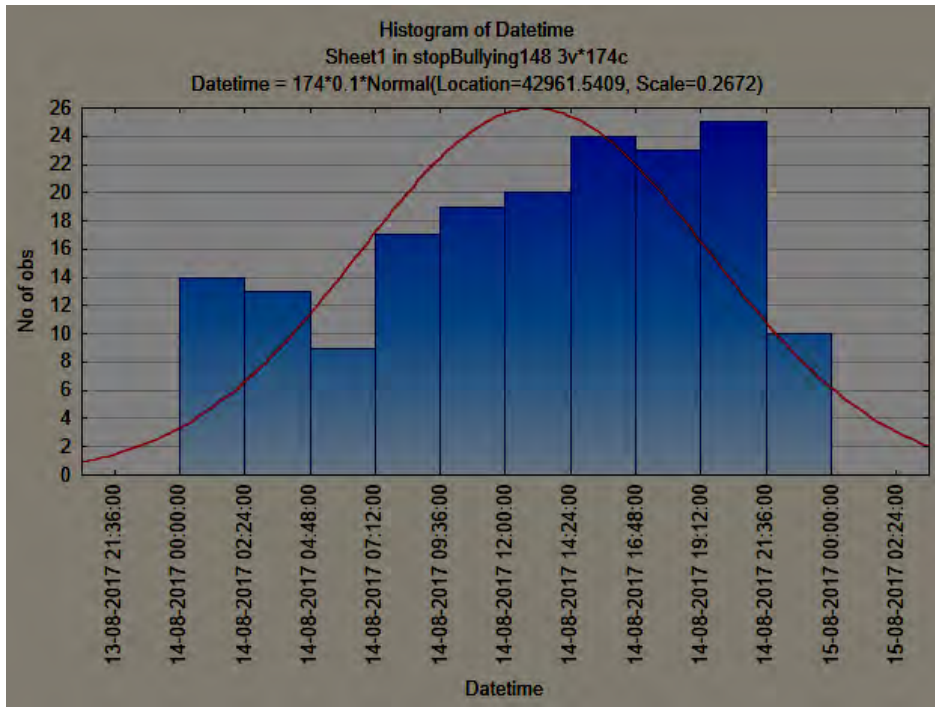


Fig: Frequency of tweets containing #stopbullying hashtag on 14-08-2017

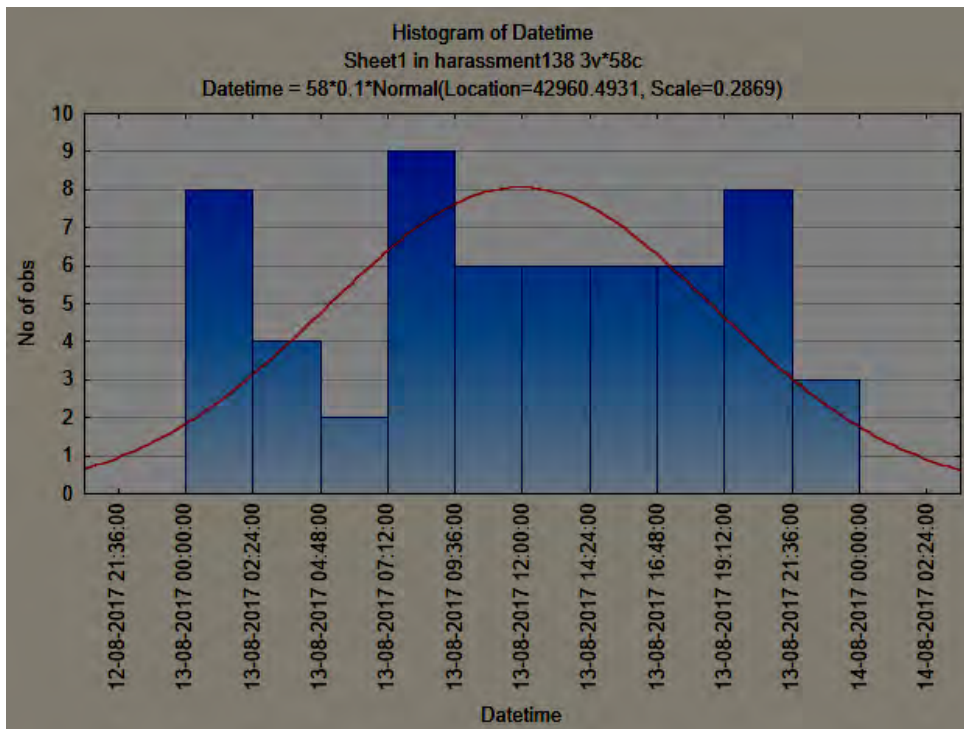


Fig: Frequency of tweets containing #harassment hashtag on 13-08-2017

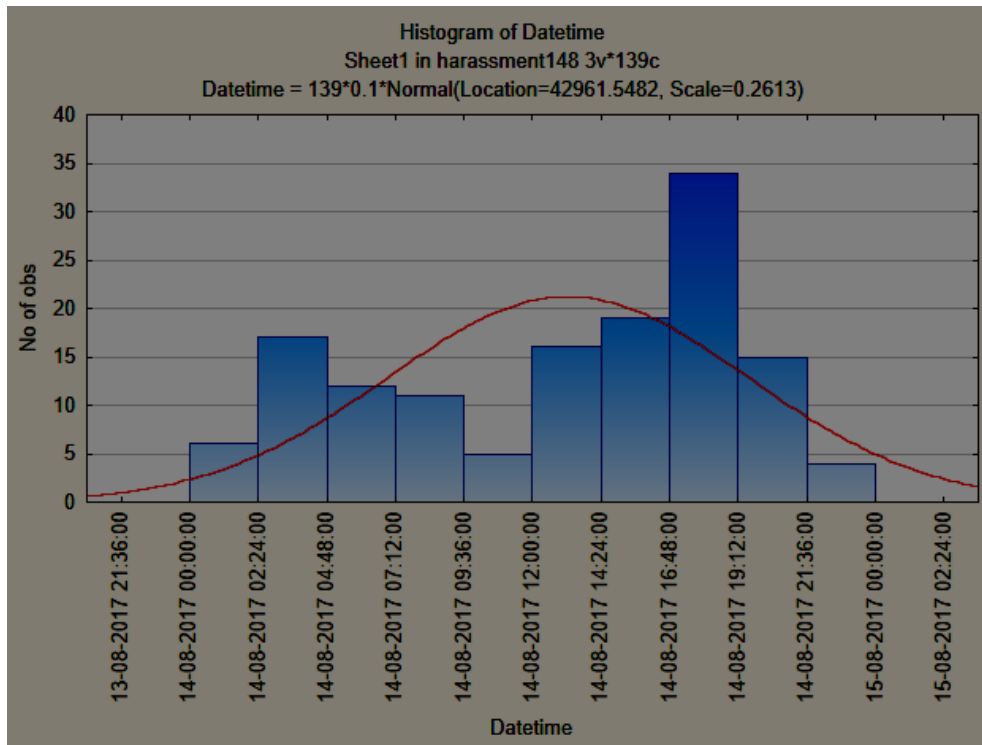


Fig: Frequency of tweets containing #harassment hashtag on 14-08-2017

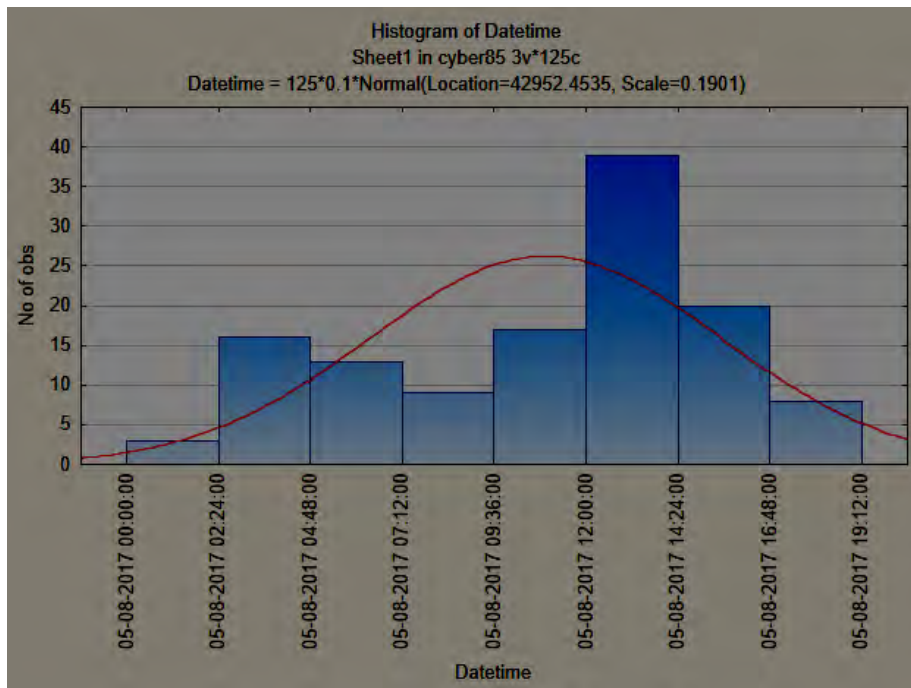


Fig: Frequency of tweets containing #cyberbullying hashtag on 05-08-2017

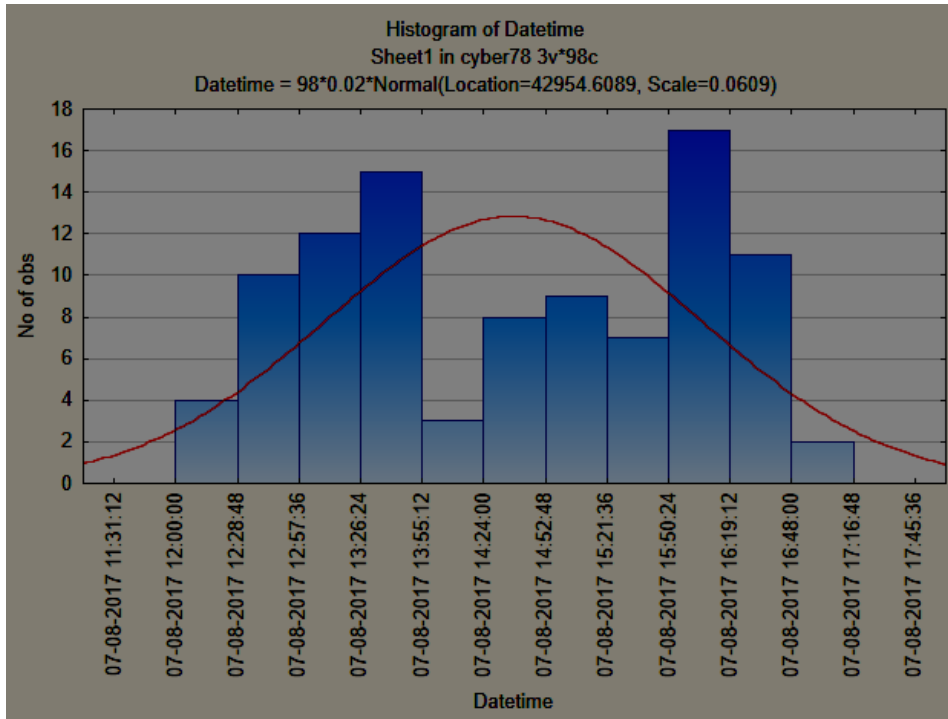


Fig: Frequency of tweets containing #cyberbullying hashtag on 07-08-2017

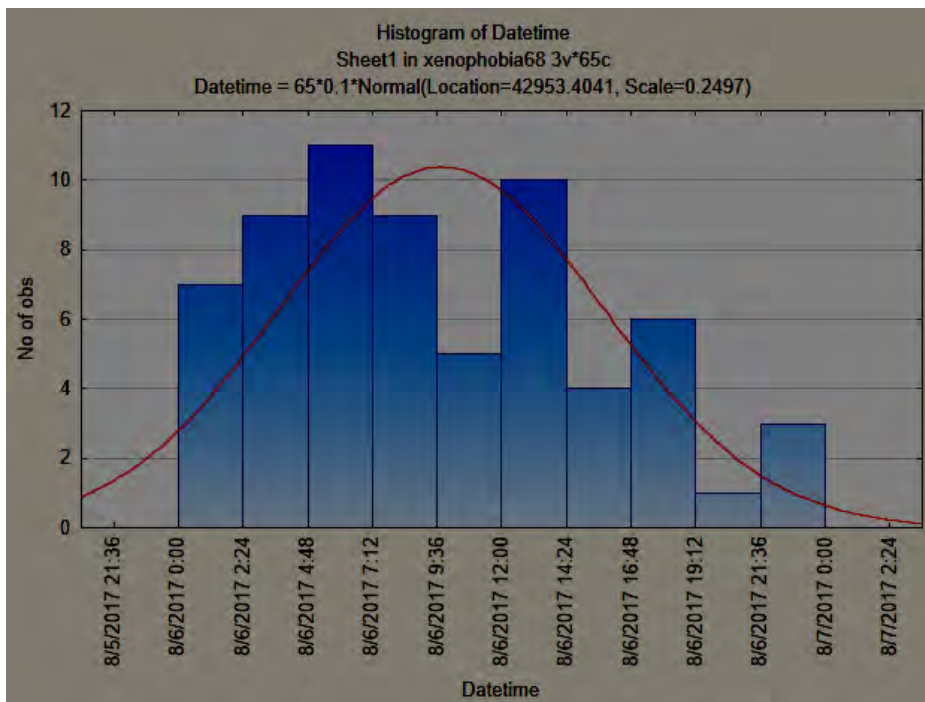


Fig: Frequency of tweets containing #xenophobia hashtag on 06-08-2017

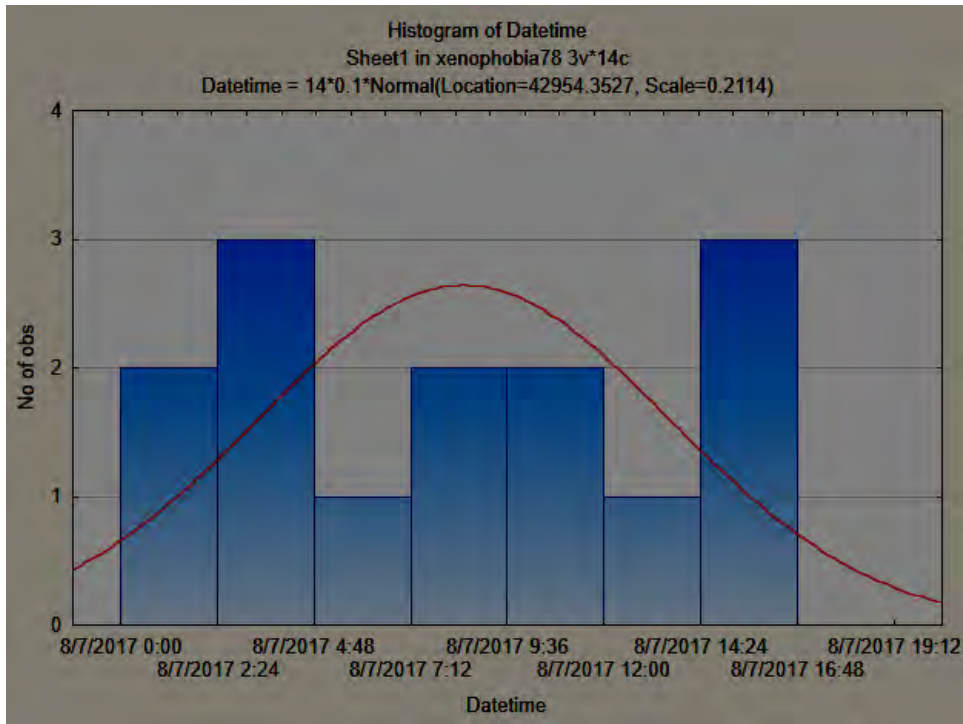


Fig: Frequency of tweets containing #xenophobia hashtag on 07-08-2017

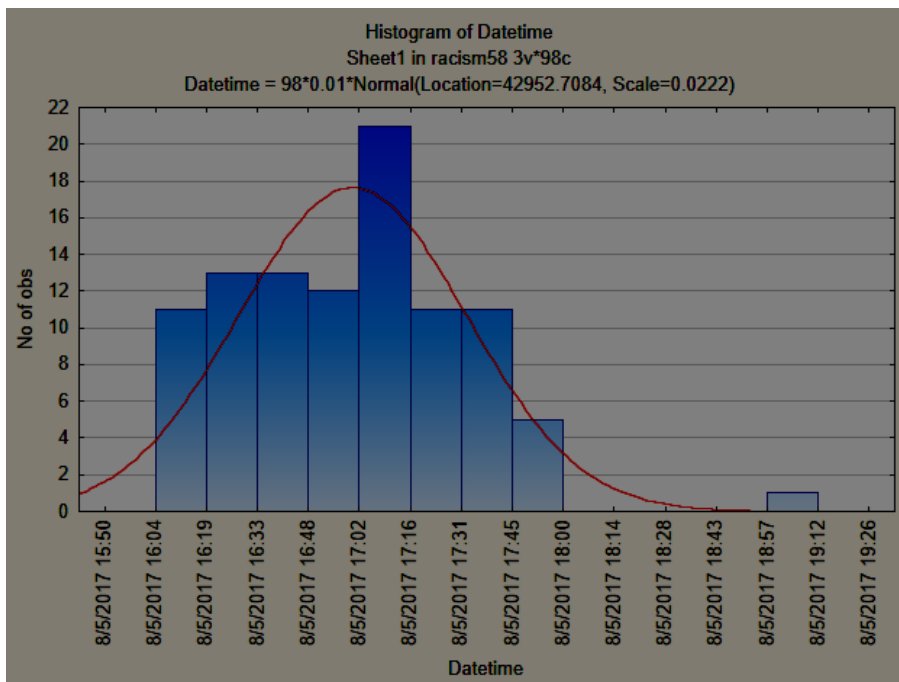


Fig: Frequency of tweets containing #racism hashtag on 05-08-2017

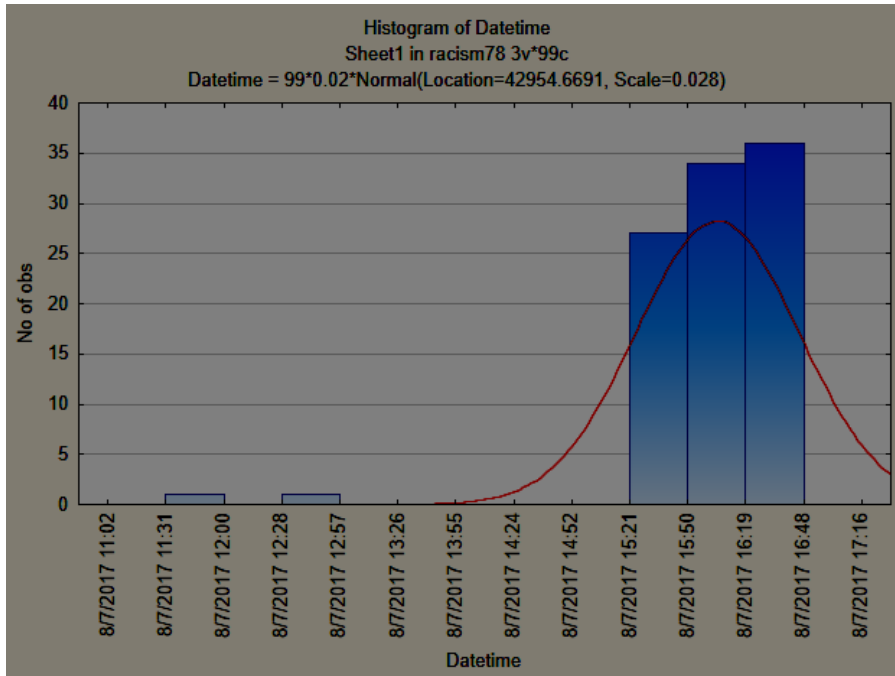


Fig: Frequency of tweets containing #racism hashtag on 07-08-2017

4.2.7 Frequency of #sexism throughout the day

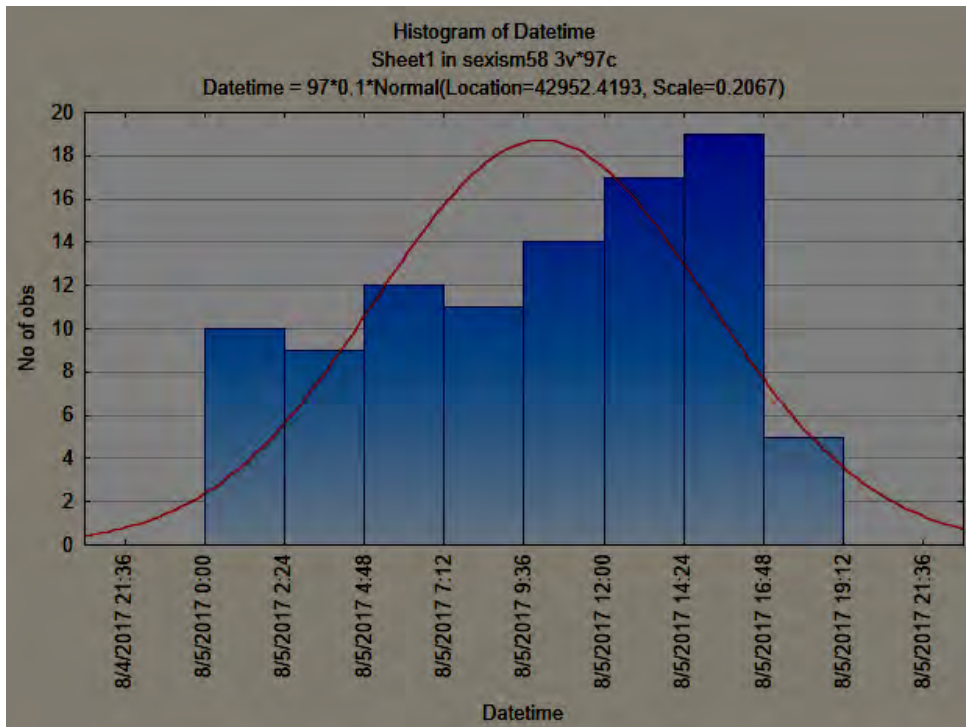


Fig: Frequency of tweets containing #sexism hashtag on 05-08-2017

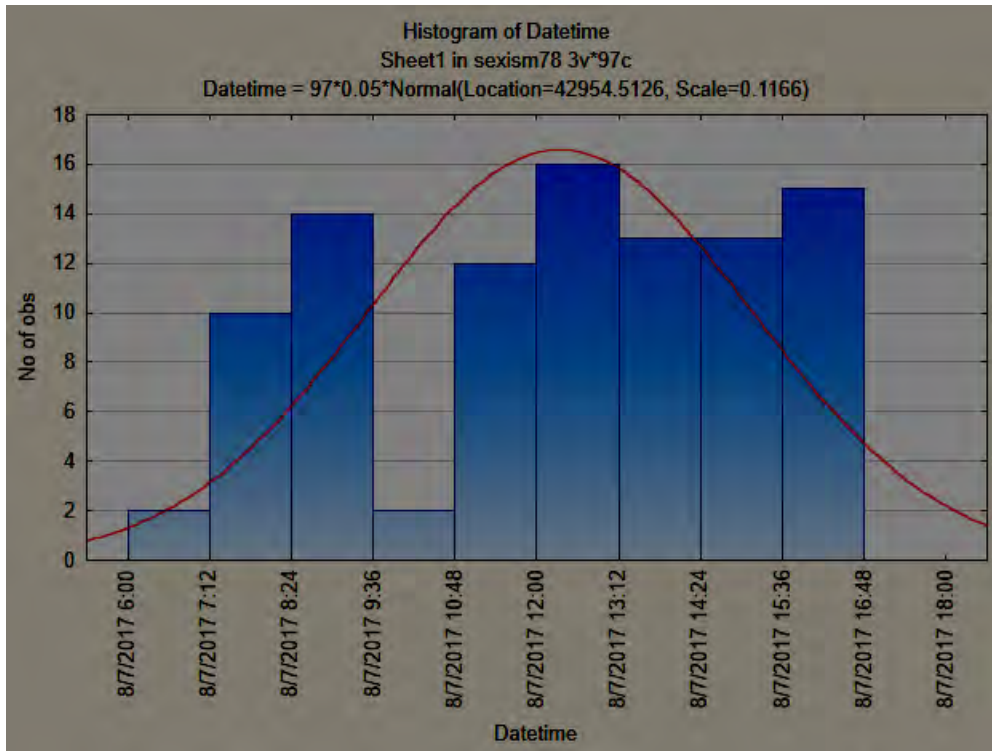


Fig: Frequency of tweets containing #sexism hashtag on 07-08-2017

4.3 Frequency of hashtags for consecutive days

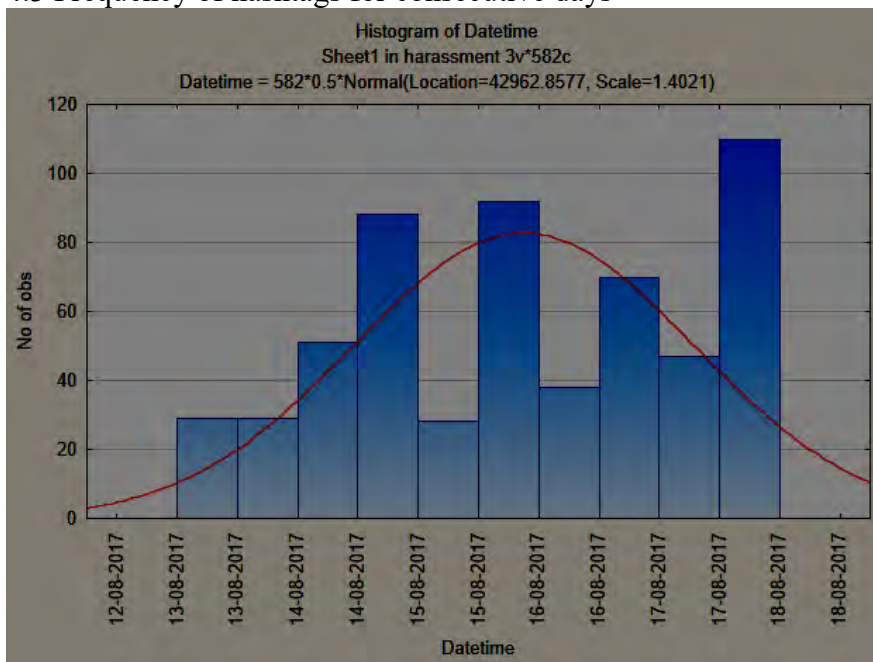


Fig: Frequency of tweets containing the hashtag '#harassment' from 13-08-2017 till 17-8-2017

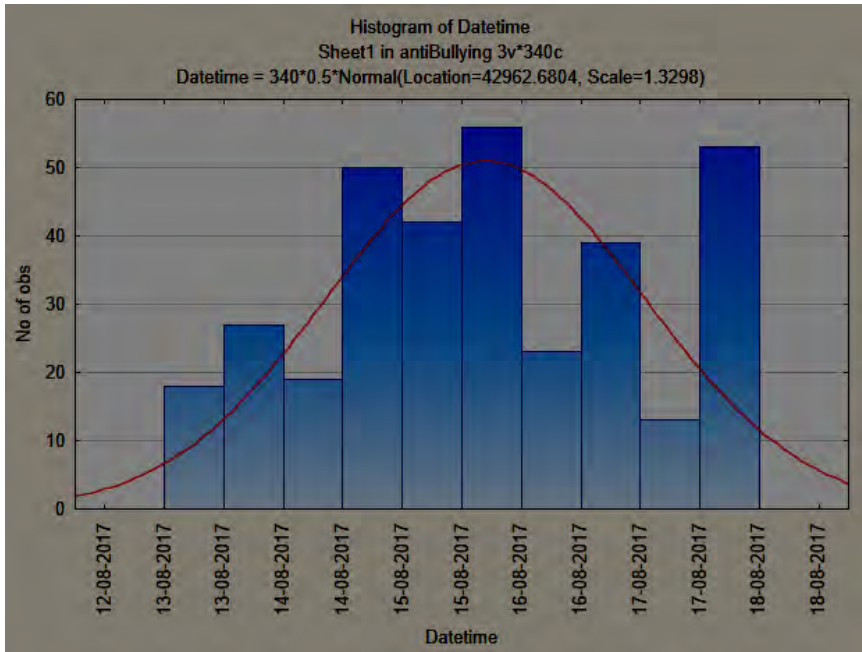


Fig: Frequency of tweets containing the hashtag '#antibullying' from 13-08-2017 till 17-8-2017

4.4 Comparison on retweets and other tweets

We compared the number of retweets with the number of other tweets, which include original tweets and replies, containing different hashtags. We had different data sets for different tweets and derived graphs from each of them. One observation here was that tweets containing the hashtag '#stopbullying' were overall retweeted the most, in comparison to other hashtags. One explanation here could be that users were retweeting stories and accounts of other people being bullied and reposting them with the hashtag '#stopbullying' as personal comments or opinions on those tweets. Additionally, among the extracted tweets consisting of '#harassment', with the exception of one day (14-08-2017), the number of retweets were overlapped by the number of original tweets or replies.

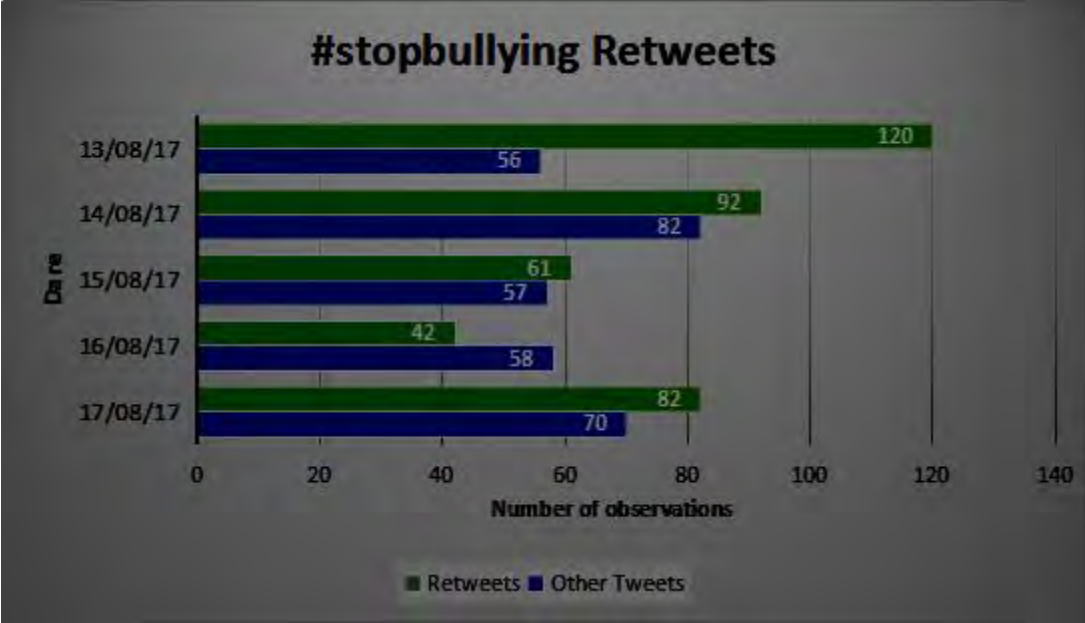


Fig: Comparison of '#stopbullying' tweets

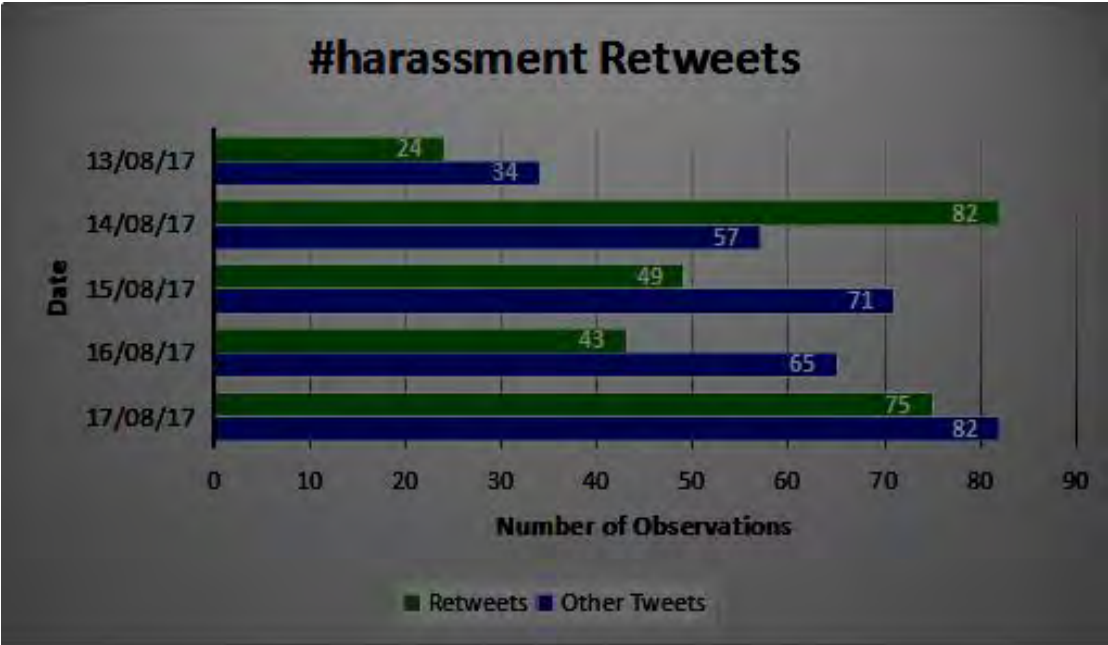


Fig: Comparison of '#harassment' tweets

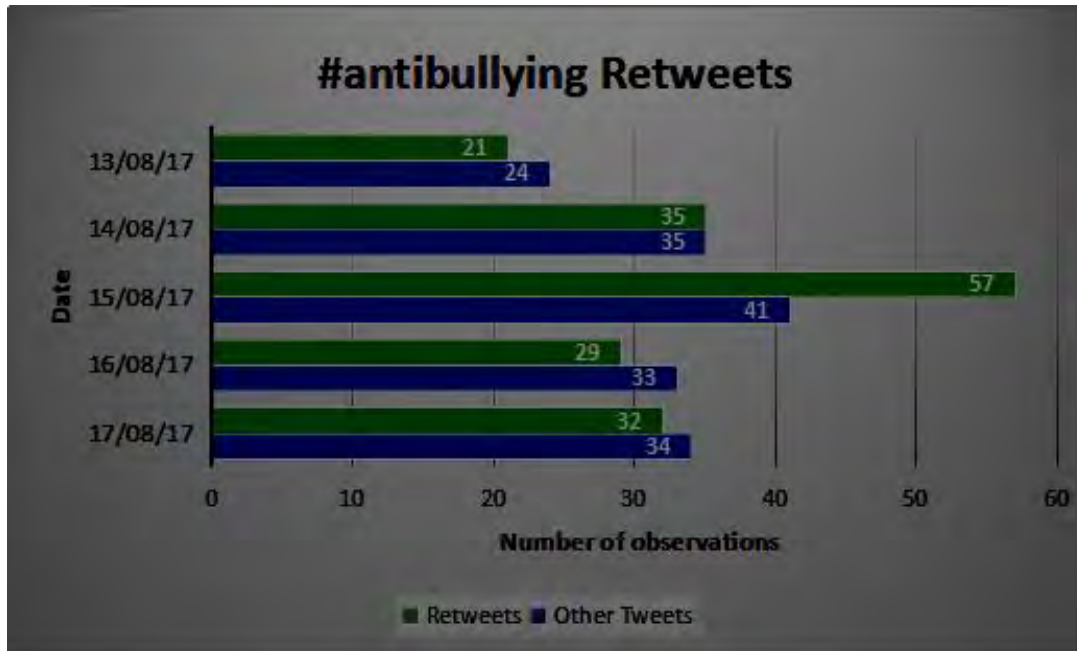


Fig: Comparison of ‘#antibullying’ tweets

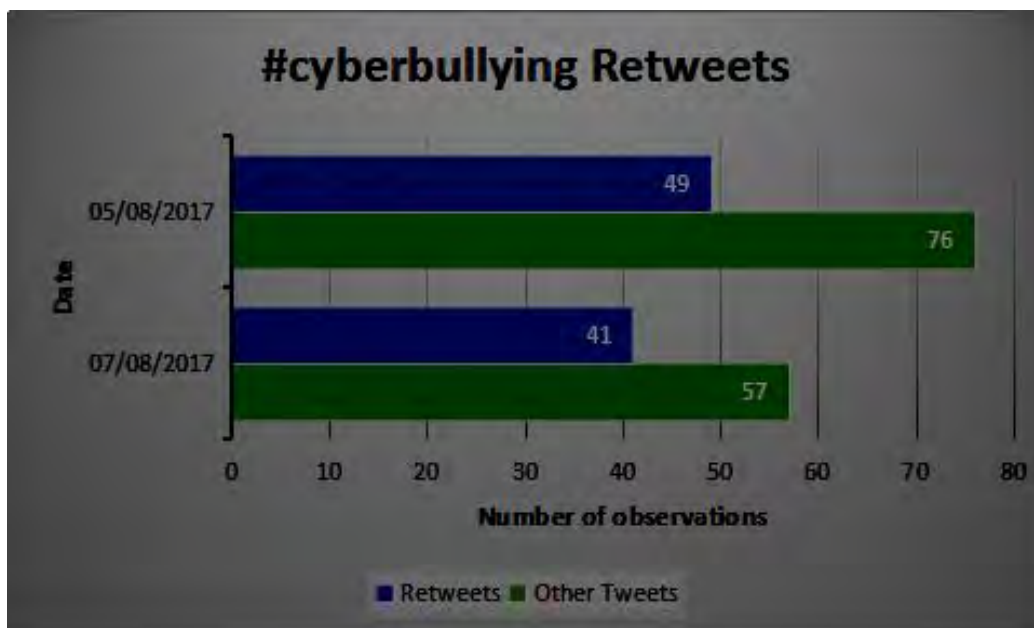


Fig: Comparison of ‘#cyberbullying’ tweets

4.5 Generation of graphs to show different use of hashtags in different locations

We have worked with 8 particular hashtags, and these are: #antibullying, #bullied, #cyberbullying, #harassment #racism, #sexism, #stopbullying and #xenophobia. For some countries, the location

shows city wise, not country wise (e.g. in US the location would show for particular cities such as Denver, LA). The text mining tool allows central time/pacific time indexing and auto time zone categorization during the text mining algorithm. The content has been used to generate graphs using the mined data to find specific hashtags in the specified locations. The next section has the screenshots of the 8 graphs and count data set is attached with it.

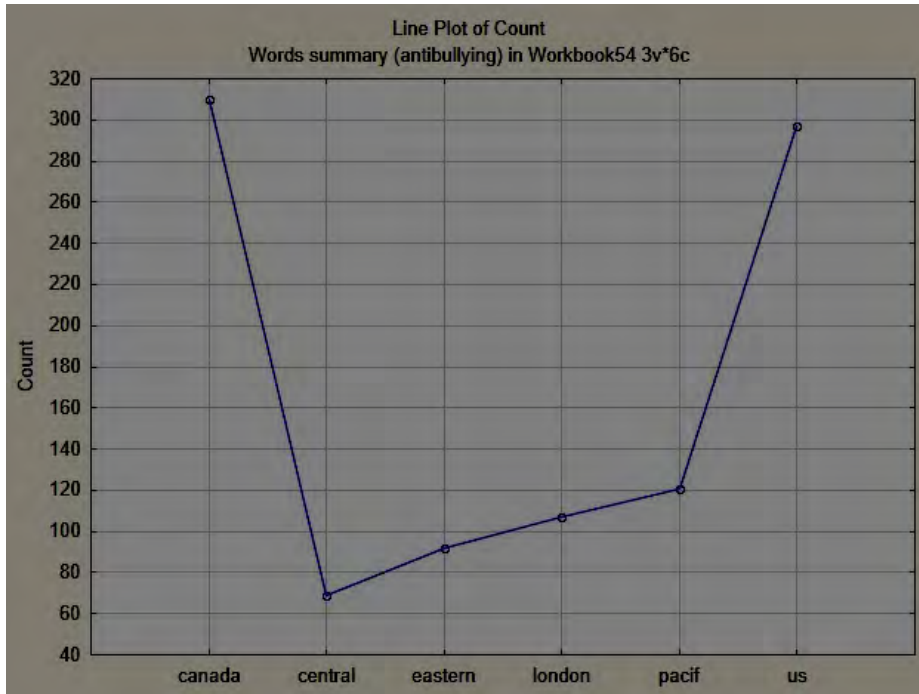


Fig: Frequency of #antibullying across different locations

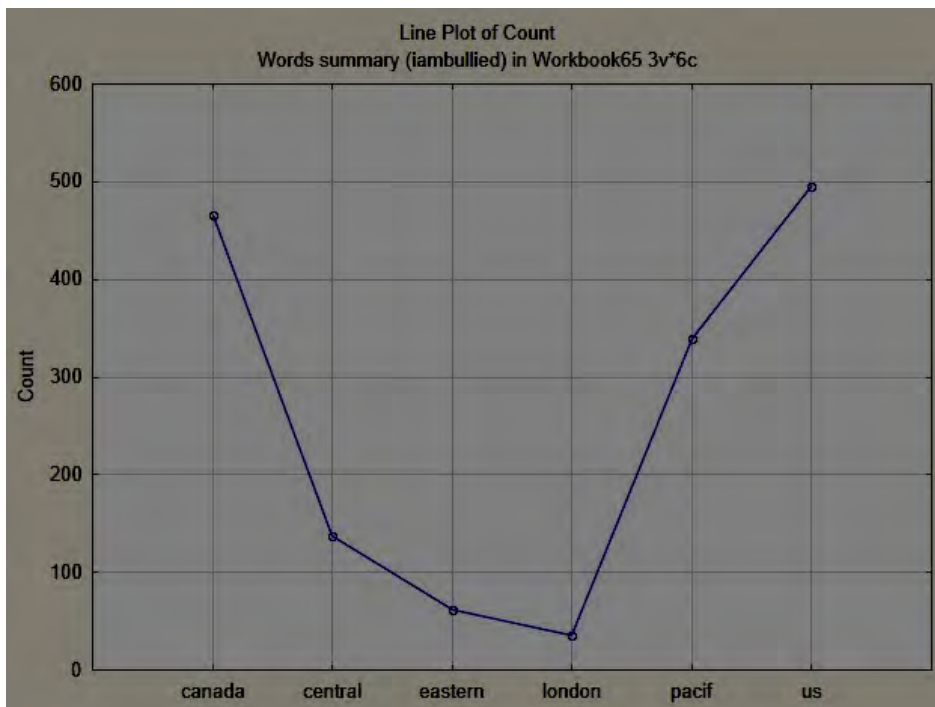


Fig: Frequency of #iambullied across different locations

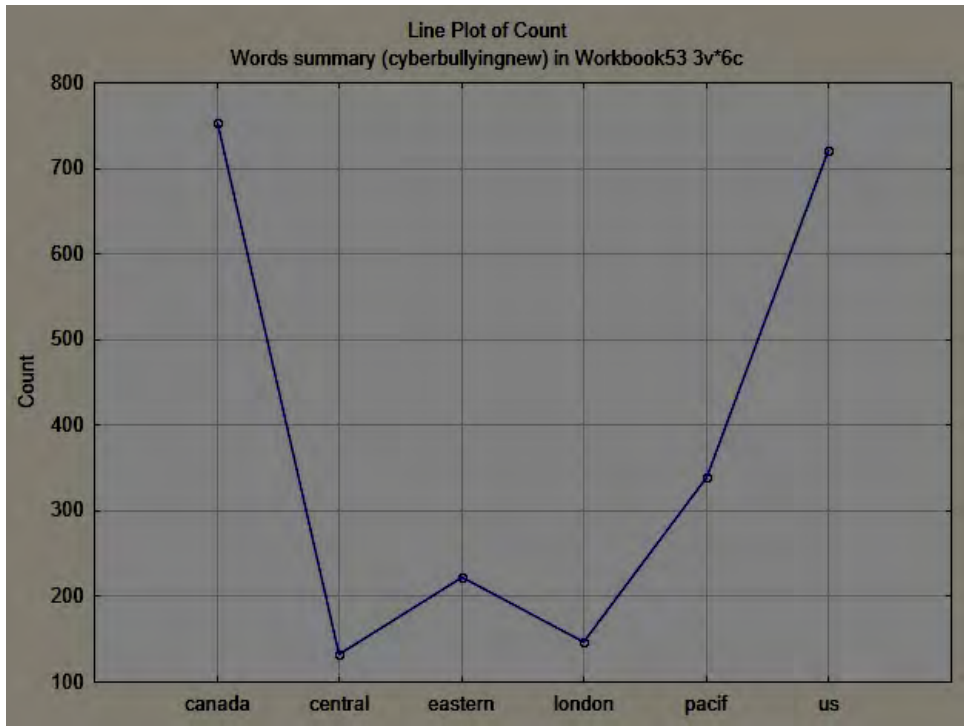


Fig: Frequency of #cyberbullying across different locations

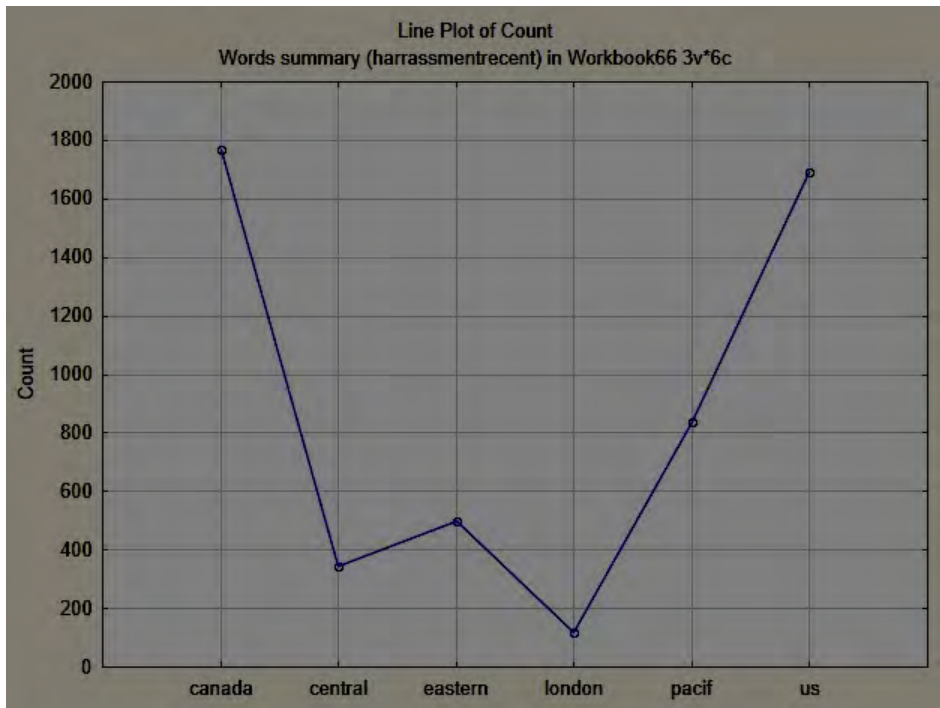


Fig: Frequency of #harrassment across different locations

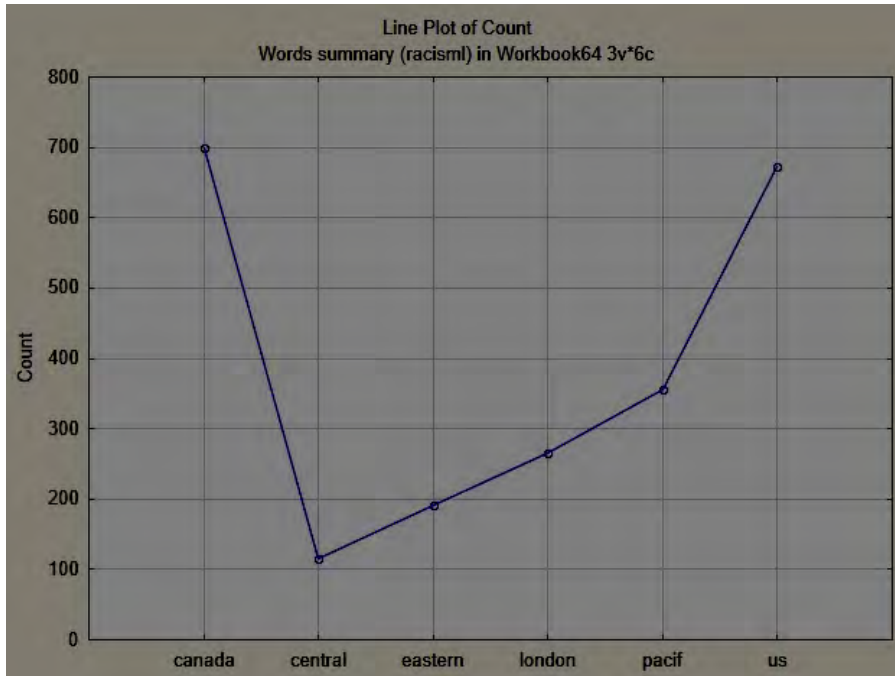


Fig: Frequency of #racism across different locations

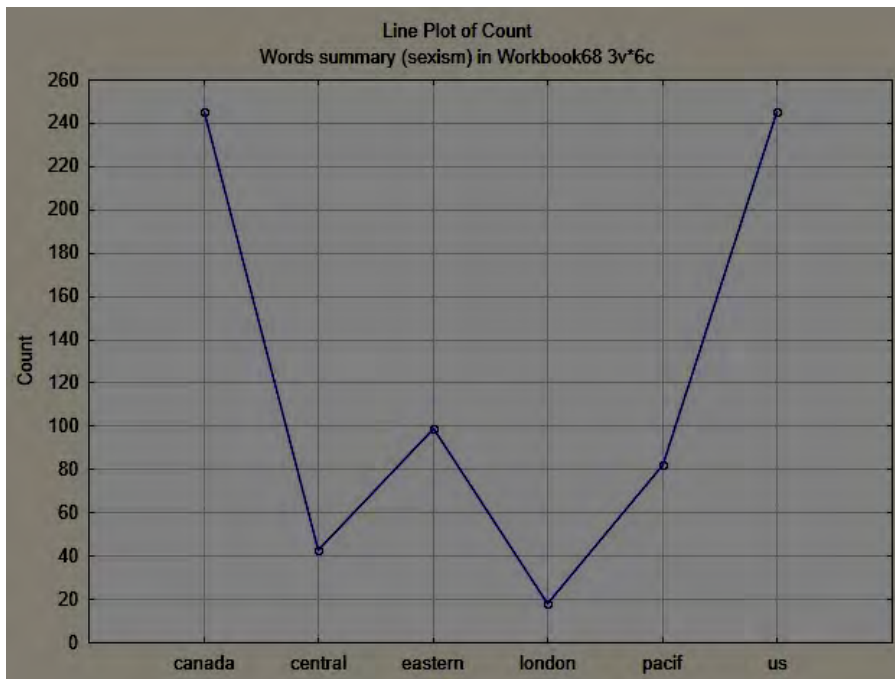


Fig: Frequency of #sexism across different locations

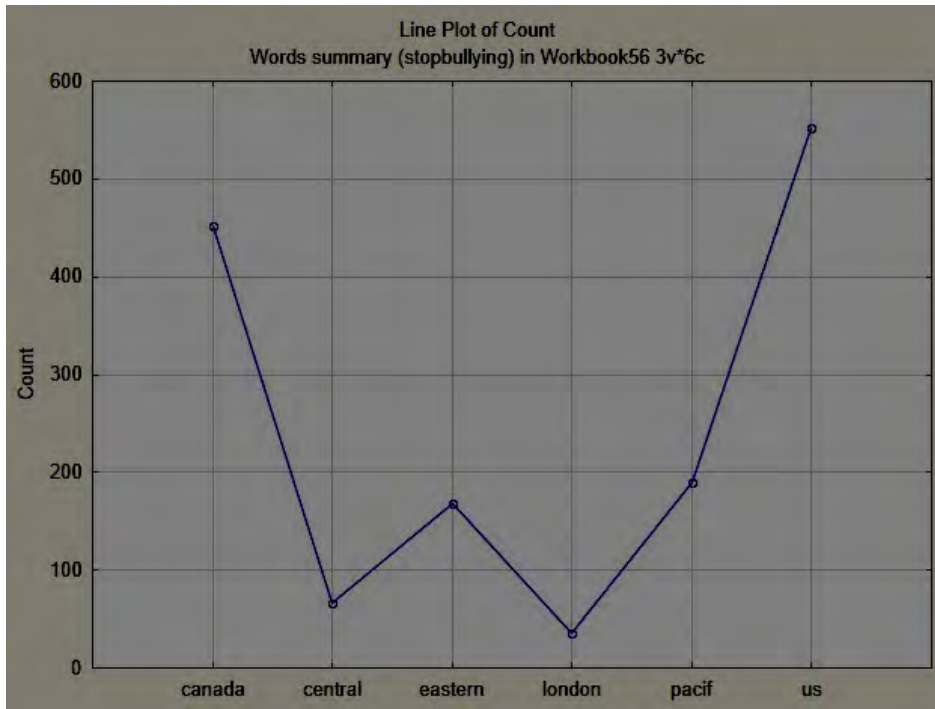


Fig: Frequency of #stopbullying across different locations

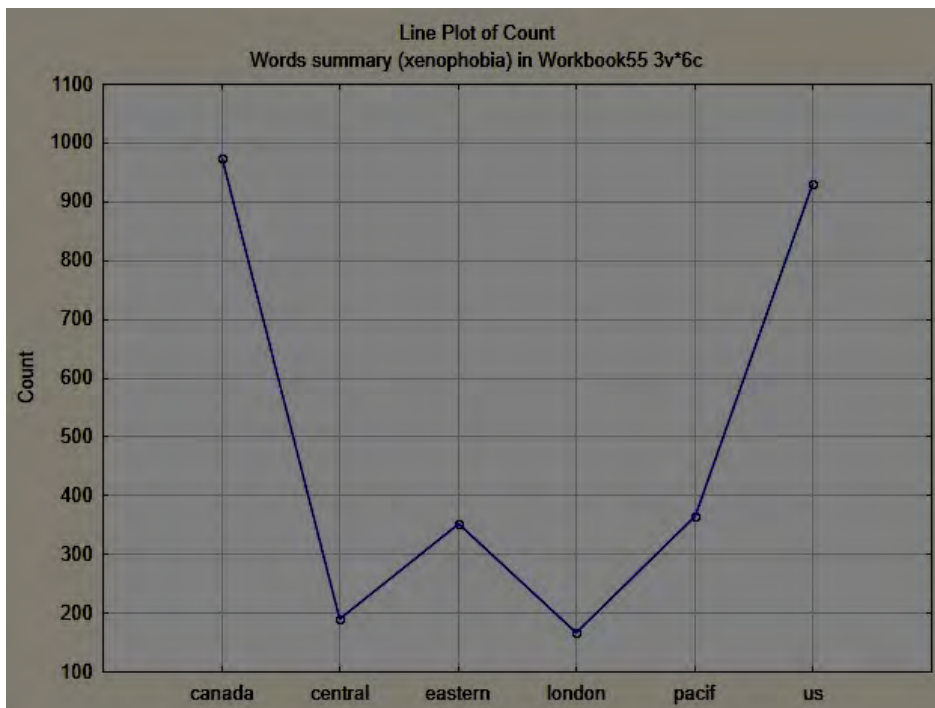


Fig: Frequency of #xenophobia across different locations

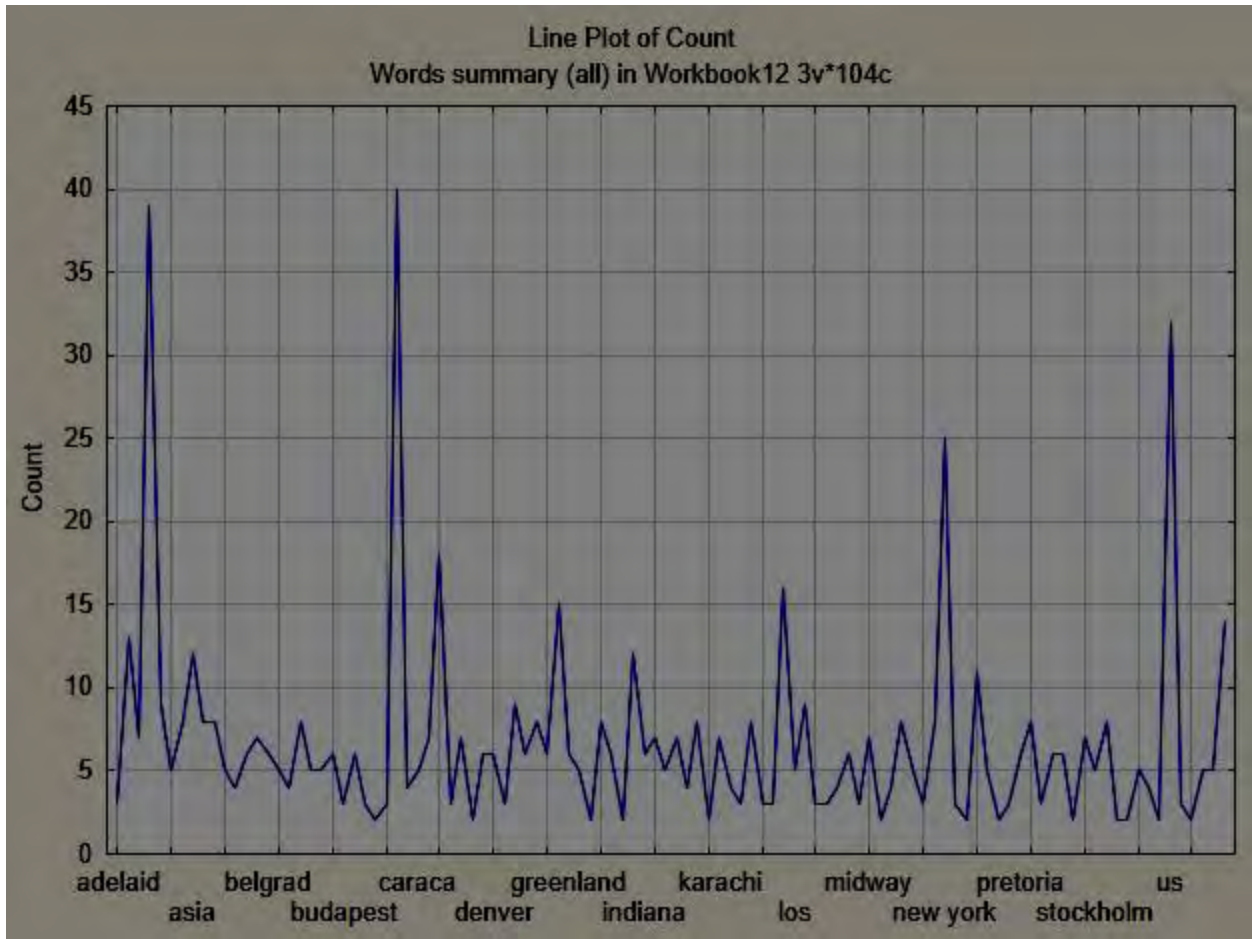


Fig: Frequency of all the hashtags with respect to all locations

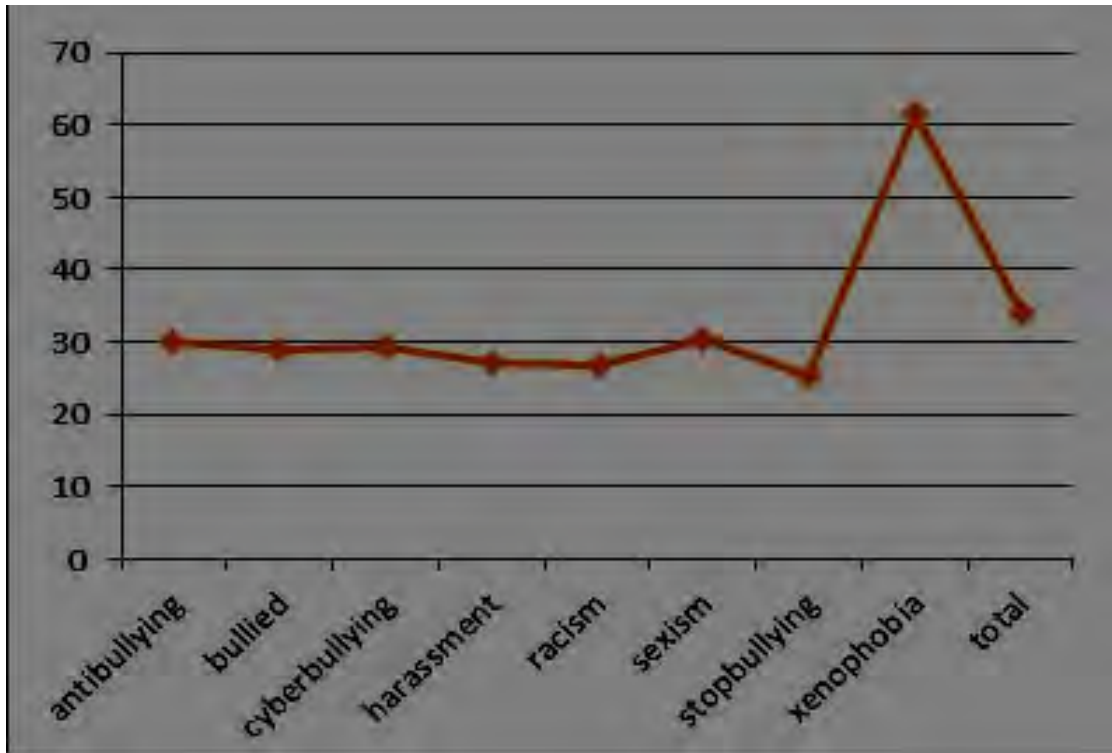


Fig: In this graph, the accuracy percentage of the location analysis for each hashtag. We can also see the total accuracy rate of the location analysis for all the hashtags, which is 34%.

If we observe the above 8 graphs, we can notice a similarity. For every hashtag related to bullying and various forms of harassment, the highest peaks are either Canada or US. This means that bullying, racism, sexism, xenophobia etc. issues exist the most in Canada and the United States out of the locations compared.

Chapter 5

Conclusion

5.1 Conclusion

Harassment and bullying have reached fever pitch in the age of social media but in turn there has also been an increase in victims coming forward and getting their voices heard. The social media provide a platform for this. Arguably the most frequently used platform for doing that is Twitter, the website boasting over 310 million users. We hoped to find a way to source this huge data resource in a way that will help to reduce bullying, create awareness and provide support groups where necessary. The aim of our study is to go through large sets of data ('tweets') and determine how frequently different kinds of harassment are reported on social media throughout the day. We started the practical implementation of our investigation by extracting tweets from Twitter using Twitter API. The tweets obtained contained specific hashtags reporting different kinds of bullying taking place in society (e.g. #sexism). These tweets were then formatted and mined using Statistica to generate graphs in terms of location, frequency and retweets. The graphs highlighted the number of tweets on specific topics at different location, the number of tweets on the said topic at different times of the day and how many of these were retweets. Based on the assumption that the number of tweets and their location mostly reflect on the situation in that place, corrective measures and help according to that can be provided.

5.2 Limitations

This study had a certain limitations at different stages. In the extraction stage, there were many hashtags to choose from which relate to bullying. The data sets would become too large if we included tweets mentioning all the terms and hashtags related to bullying so we chose had to choose only a few of them. Another drawback here is that Twitter's REST API sometimes extracts all the tweets from certain days and one or two tweets from some others, due to which we had to discard the values of those dates. People do not always post serious tweets, some of them are posted in jest or for other reasons. This meant that not all the tweets extracted were addressing

serious incidents or opinions related to bullying. One of the main limitations we encountered was during the stage of mining data according to different locations. Comparisons could not be drawn directly between countries, rather, the information from those countries were divided into their regions. For example, in the US, the locations were divided into states, instead of the whole country overall.

5.3 Future Work

One of the main components of this study was tracking different hashtags according to locations around the world. In the future, this could be expanded to user age and gender, which would give us an even clearer picture of the extent of bullying in society and how it is shared and reported. As of now, Twitter APIs do not work on these parameters, however, in the long run, algorithms could be derived from users' tweets, photos and "Bio" fields. Data could also be aggregated for years to see when the frequency of tweets related to bullying were being posted and to check if any spikes in data corresponded to public events, like a terror attack.

References

6.1 List of References

1. Go, Bhayani, Huang, Twitter Sentiment Classification using Distant Supervision <http://www.stanford.edu/~alecmgo/cs224n/sigproc-sp.pdf>
2. Pennacchiotti, M. & Popescu, A.-M. (2011). A Machine Learning Approach to Twitter User Classification. ICWSM, 11, 281--288.
3. Go, A., Bhayani, R. & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. Processing, 1--6.
4. Pak, A. & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2886/3262>
5. Jansen, B. J., Zhang, M., Sobel, K. & Chowdury, A. (2009). Twitter Power: Tweets as Electronic Word of Mouth. Journal of the American Society for Information Science and Technology, 60(11), 2169–2188.doi: 10.1002/asi.21149
6. Xu, Jun-Ming, Zhu, Xiaojin & Bellmore, Amy (2012). Fast Learning for Sentiment Analysis on Bullying, Retrieved from <http://pages.cs.wisc.edu/~jerryzhu/pub/wisdom12.pdf>
7. Todd Wasserman (Dec 3, 2012). "[McDonald's Releases First TV Ad With Twitter Hashtag](#)". Mashable
8. Cortis, Keith (2015). Analysis of cyberbullying tweets in trending world events
9. "[What is hashtag?](#)", Mashable, 8 October 2013
10. Kricfalusi (March, 2017). "The twitter hashtag: What is it and how do you use it?".Tech for Luddites
11. Goyal, Diwakar (2011). Data Mining and Analysis on Twitter
12. Kumar, Morstatter, Liu (2013). Twitter Data Analytics
13. "[Big Data Definition](#)". MIKE2.0. Retrieved 9 March 2013.
14. Auth Flow. Retrieved from <https://dev.twitter.com/>

15. statsofsa(June 10 2014) Retrieved from <https://statisticasoftware.wordpress.com/>
16. Retrieved from <https://www.predictiveanalyticstoday.com/statistica/>
17. Wass, John (July 12, 2011). “STATISTICA 10: The power of Statistics and Data Mining Simplified”.
18. Retrieved from
<http://documentation.statsoft.com/STATISTICAHelp.aspx?path=TextMiner/TextMiner/TextMiningResultsDialog>
19. Manning, Schütze (2012). Foundations of Statistical Natural Language Processing
20. Retrieved from <https://www.linguamatics.com/what-is-text-mining>
21. Retrieved from http://user.engineering.uiowa.edu/~ie_230/lecture/Text_Mining.pdf
22. Nisbet, Elder, Miner (2009). Handbook of Statistical Analysis & Data Mining Applications
23. Abbasi A, Rashidi TH, Maghrebi M, Waller ST, (2015). 'Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time', in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*
24. Nasukawa, Yi (2003). Sentiment Analysis : Capturing Favorability Using Natural Language Processing Definition of Sentiment Expressions, K-CAP '03 Proceedings of the 2nd
25. Angela J. Calvin, Amy Bellmore, Jun-Ming Xu & Xiaojin Zhu (2015). #bully: Uses of Hashtags in Posts About Bullying on Twitter, *Journal of School Violence*, 14:1, 133-153, DOI: 10.1080/15388220.2014.966828