Project Proposal

Identifying Major Social Media Events from Twitter



|  |  |
| --- | --- |
| Name: | Supervisor: |
| University ID: |  |
| Word Count: 3885 (Without References) |  |
| Date: |  |

Table of Contents

[Table of Figures 2](#_Toc68949269)

[Abstract 3](#_Toc68949270)

[Introduction 3](#_Toc68949271)

[Background 3](#_Toc68949272)

[Objectives/Aim 4](#_Toc68949273)

[Ethical Considerations 4](#_Toc68949274)

[Content consistence 5](#_Toc68949275)

[Off-Twitter coordinating 5](#_Toc68949276)

[Utilizing geo-information 6](#_Toc68949277)

[Twitter passwords 7](#_Toc68949278)

[Project Timeline 7](#_Toc68949279)

[Methodology 7](#_Toc68949280)

[Literature Review 7](#_Toc68949281)

[GraphIE: A Graph-Based Framework for Information Extraction 8](#_Toc68949282)

[Information extraction from text messages using data mining techniques 10](#_Toc68949283)

[Experimentation 11](#_Toc68949284)

[Dataset 11](#_Toc68949285)

[Machine Learning 12](#_Toc68949286)

[Project Evaluation 12](#_Toc68949287)

[Conclusion 12](#_Toc68949288)

[References 13](#_Toc68949289)

# Table of Figures

[Figure 1 - Gantt chart showing project timeline 7](#_Toc68877105)

[Figure 2 - Literature Review Methodology 7](#_Toc68877106)

[Figure 3 - Architecture and Framework for GraphIE 8](#_Toc68877107)

[Figure 4 - Results 9](#_Toc68877108)

[Figure 5 - Statistics for EDUCATION and JOB datasets 10](#_Toc68877109)

[Figure 6 – Results 10](#_Toc68877110)

Abstract

Social Media is slowly evolving. Evolving into a tool with multiple uses rather than just a platform for connecting with friends and family. People on social media have started realizing the prowess of social media over the years. Whether it be online marketing, fame gathering, news networks, forums or anything else from any domain, social media is now becoming the go-to for all.

This revelation has brought about the radical increase in data created on social media. The amount of data created makes it difficult for people to search or see the highlights or popular events on social media. This research project is an experimental approach towards handling this problem using Machine Learning. This project provides experimental data and analysis on the performance on Machine Learning for this problem as an output. The problem this research will be solving is the identification and extraction of major events from twitter.

**Keywords:** Machine Learning, Experimental Data and Analysis.

Introduction

Online media innovation has guided a period empowering remarkable freedom for people to interface with each other (Martinez, Tsou and Spitzberg, 2019). From amusement to legislative issues, news and discussions on arrangements of different sorts, web-based media is taking over quickly. This has likewise helped the creation of information dramatically. Analysts have now begun exploring for approaches to extricate data without any problem.

Data mining from web-based media content has as of late become a functioning examination point, following early trials that demonstrated this type to be very trying for best-in-class calculations (Bontcheva and Derczynski, 2016). For finding arising information, there is today another and very incredible source: socially created content. Interpersonal organizations are gigantic (1.8 billion of clients) and new (their substance is created while occasions happen) (Brambilla et al., 2017).

This section will continue to discuss the background, aim, the research questions, ethical considerations and the project timeline in that order. The methodology section after this will explain the processes involved in Literature Review and Experimentation. This will be followed by a conclusion, references, appendix and glossary.

## Background

Twitter reigns as one of the most popular social media on the internet today. Since its humble beginnings in 2006, twitter has become the face of social and political fusion. With big celebrities and political figures sharing their thoughts and pens on twitter, there are many posts that can be considered as major events on twitter.

Naturally, there are many people that share their own opinions on these major posts by quoting or retweeting them thereby causing a chaotic net of interconnected circles of retweets. Navigating through all of this to get to the original post can be a little hectic sometimes. This project can help people to quickly find hot news they’re looking for.

## Objectives/Aim

The aim of the project as described is to create a machine learning algorithm/ensemble that can identify and extract major social events on twitter. To better improve the quality of the research experiments, the research is focused on the current affairs and news domain for identification of major events. This will be done though an experiment following these steps:

* Dataset finalization and comprehension: The dataset is the most important and crucial part of any machine learning based research project. The finalizing and understanding of a dataset for this research will boost the progress of this project by providing training material for the ML algorithm.
* Machine Learning Algorithm Creation and Finalization: The machine learning algorithm is the next most crucial part of this research. The algorithm is responsible for digesting the above dataset and providing us the results we require, i.e., the identification and extraction of major events on social media.
* Empirical Analysis of Algorithm Performance: Statistical analysis of the algorithm will provide the reader with performance metrics and graphs that convey how good the performance of the algorithm was.

## Ethical Considerations

Twitter has an extensive policy for the usage of data from twitter. The **Consent & permissions** section on the twitter guidelines present the concerns as follows:

Making any moves for their sake. This incorporates (yet isn't restricted to):

* Posting substance to Twitter
* Following/unfollowing accounts
* Modifying profile or record information
* Starting a Periscope Broadcast
* Adding hashtags or some other substance to Tweets
* Republishing content got to by infers by some other means than methods for the Twitter API or other Twitter gadgets
* Using someone's Twitter Content to propel a thing or organization
* Storing non-public substance like Direct Messages (DMs), or some other private or restricted information
* Sharing or appropriating guaranteed substance, or some other private or privileged information

In the event that your administration permits individuals to present substance on Twitter you should do the accompanying prior to distributing:

* Show precisely what will be distributed
* Make it clear to individuals utilizing your administration what geo data (assuming any) will be added to the substance

On the off chance that your administration permits individuals to present substance on both your administration and Twitter, you should do the accompanying prior to distributing:

* Obtain consent to post the substance
* Explain where you will post the substance
* You should regard the secured and impeded status of all Twitter Content. You may not serve content got utilizing one individual's confirmation token to an alternate individual who isn't approved to see that content.

Ensured accounts: A secured record's substance is simply accessible to individuals who have been affirmed by the proprietor to follow that account. Along these lines, on the off chance that you run a help that gets to secured accounts, you may just do as such to serve such substance to the particular individuals with consent to see that content.

Blocked accounts: People on Twitter can obstruct admittance to their records under any condition they pick. Coexisting data got from tokens (or some other API-based activity) to sidestep this decision isn't allowed.

As Direct Messages (DMs) are non-public in nature, benefits that give DM highlights should find additional ways to defend individual security. You may not serve DM substance to individuals who are not approved to see that content. On the off chance that your administration gives DM usefulness you should likewise:

Tell individuals on the off chance that you send read receipt occasions for DMs. You can do this by giving a notification straightforwardly in your administration, or by showing read receipts from different members in a discussion.

### Content consistence

In the event that you store Twitter Content disconnected, you should stay up with the latest with the present status of that content on Twitter. In particular, you should erase or adjust any substance you have on the off chance that it is erased or changed on Twitter. This should be done when sensibly conceivable, or inside 24 hours in the wake of accepting a solicitation to do as such by Twitter or the pertinent Twitter account proprietor, or as in any case needed by your concurrence with Twitter or appropriate law. This should be done except if in any case denied by law, and really at that time with the express composed consent of Twitter. Adjusted substance can take different structures. This incorporates (however isn't restricted to):

* Content that has been made private or gained guaranteed status
* Content that has been suspended from the stage
* Content that has had geotags taken out from it
* Content that has been held or taken out from Twitter

### Off-Twitter coordinating

The conditions under which you may coordinate with an individual on Twitter to data got or put away off-Twitter are limited. Off-Twitter coordinating includes partner Twitter Content, including a Twitter @handle or client ID, with an individual, family, gadget, program, or other off-Twitter identifier. You may possibly do this in the event that you have express pick in assent from the individual prior to making the affiliation, or as depicted beneath.

In circumstances in which you don't have an individual's express, select in agree to interface their Twitter character to an off-Twitter identifier, we necessitate that any association you draw be founded uniquely on data that somebody would sensibly hope to be utilized for that reason. Also, missing an individual's express pick in assent you may just endeavor to coordinate with your records about somebody to a Twitter character dependent on:

* Information gave straightforwardly to you by the individual. Note that records about people with whom you have no earlier relationship, including information about people acquired from outsiders, don't satisfy this guideline; as well as
* Public information. "Public information" in this setting alludes to:
* Information about an individual that you got from a public, by and large accessible asset (like a catalog of individuals from an expert affiliation)
* Information on Twitter about an individual that is openly accessible, including:
  + Tweets
  + Profile data, including a record bio and openly expressed area
  + Display name and @handle

Your security strategy

You should show your administration's security strategy to individuals before they are allowed to download, introduce, or join to your administration. It should uncover in any event the accompanying data:

* The data that you gather from individuals who utilize your administration
* How you use and offer that data (counting with Twitter)
* How individuals can reach you with requests and demands in regards to their data

Your security strategy should be reliable with all material laws, and be no less defensive of individuals than Twitter's Privacy Policy and the protection strategy of our different administrations and corporate members. You should stop your admittance to the Twitter API and the utilization of all Twitter Content in the event that you can't agree with your or potentially Twitter's Privacy Policy.

## Project Timeline

The need to divide the project into smaller modules and submodules is of critical importance. Assigning those deadlines and performing them one by one keeps the developer focused and concentrated on efficient work. This is known as project timeline. Gantt charts are the standard when it comes to representing project timelines for projects. The project timeline for this project is as follows:

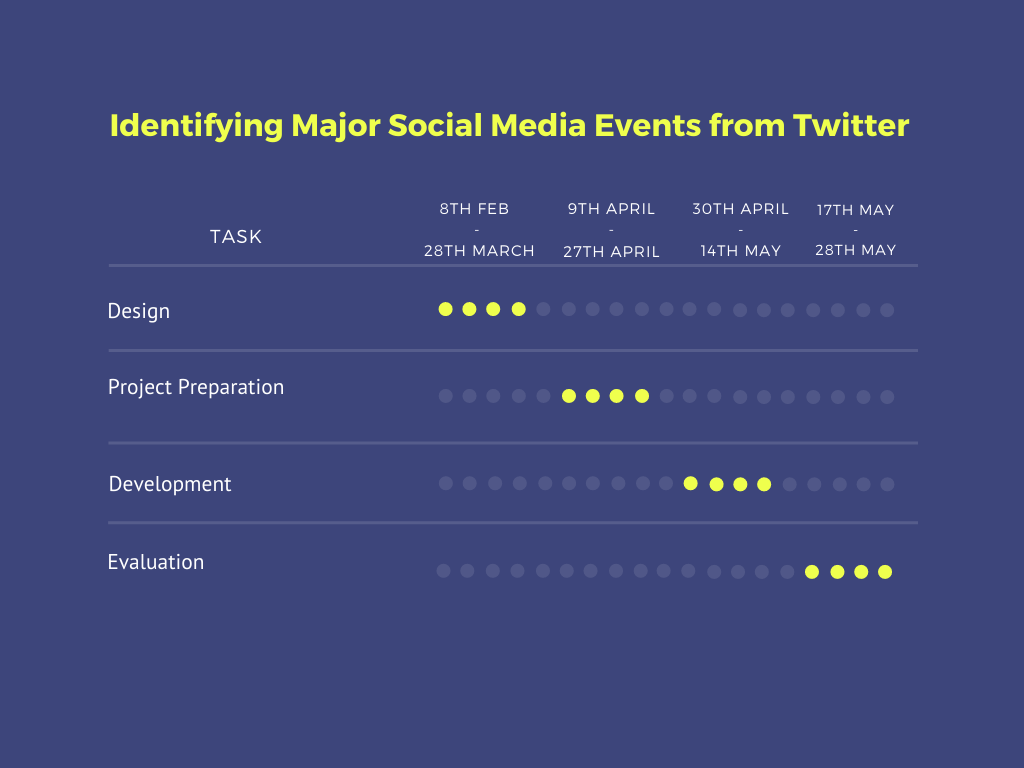


Figure 1 - Gantt chart showing project timeline

# Methodology

The different parts that are relied upon to procedurally arrange the task towards satisfaction make up the Methodology. Execution of this method gives a ton of presumes that makes movement more smoothed out and centered towards the fundamental goals.

## Literature Review

The assessment of online media was several years earlier. The investigation on this region started by then. From that time, there has been a huge load of work done in this field. A strong and properly performed LR can transform into the endorsement point for the reliability of the undertaking. In view of its importance, a specific methodology has been allotted to play out this task successfully, as displayed in *Figure 2*.

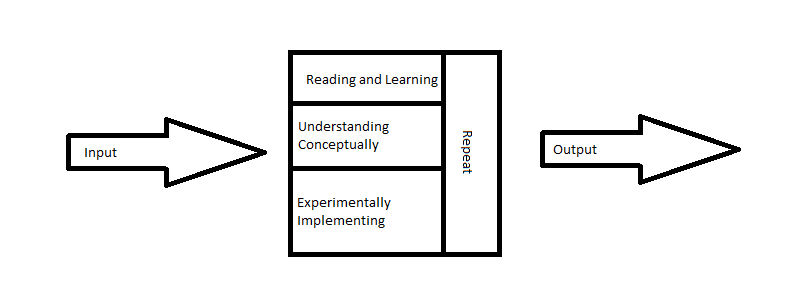


Figure 2 - Literature Review Methodology

As referred to above, work is being done on this point since several years back. In this time, a lot of review material, specifically investigation papers have amassed for anyone that necessities to do a Literature Review. However, looking over all of the papers as yet isn't pragmatic. Subsequently, a course of action for picking quality papers needs to in be place. The course of action is according to the accompanying:

* English ought to be the singular language in the papers.
* Ought to be spread in journals with a high effect factor.
* Complete and free access should be accessible for the Paper and the Journal.
* All the evaluation datasets and code for the papers should be open complimentary.
* A decade old research is not allowed.

### GraphIE: A Graph-Based Framework for Information Extraction

In this paper, GraphIE is presented. It is a system that works over a chart addressing an expansive arrangement of conditions between literary units (i.e., words or sentences). The calculation proliferates data between associated hubs through diagram convolutions, producing a more extravagant portrayal that can be misused to improve word-level expectations.

GraphIE improves predictions by automatically learning the interactions between local and non-local dependencies in the input space. This approach integrates a graph module with the encoder-decoder architecture for sequence tagging. At the core of this model, a recurrent neural network sequentially encodes local contextual representations and then the graph module iteratively propagates information between neighboring nodes using graph convolutions. The learned representations are finally projected back to a recurrent decoder to support tagging at the word level.

#### Methodology

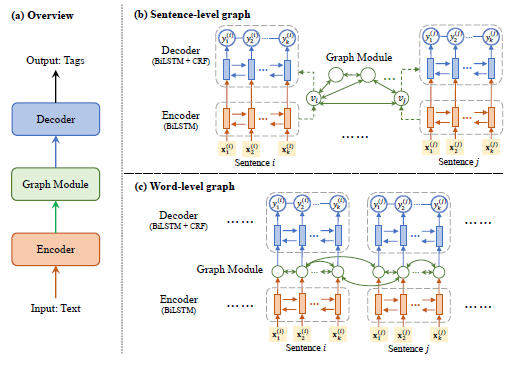


Figure 3 - Architecture and Framework for GraphIE

Three components in the GraphIE model:

1. The **encoder** generates local context-aware hidden representations for the textual unit with a Recurrent NN. The encoder RNN is defined by the following: where x denotes the input sequence, h denotes the hidden states and denotes the encoding parameters.
2. The **graph module** captures the graphical structure between textual units. The graph module is a Graph ConvNet (GCN). It consists of two parts:
   1. The first part gets the information from all nodes in the previous layer: where W is the weight for each input.
   2. The second part aggregates information from the neighbours of each node: where d is the degree of node (i.e. the number of edges connected to) and is used to normalize, ensuring that nodes with different degrees have representations of the same scale.
3. The **decoder** uses the data generated by the graph module to perform labelling at the word level. It consists of a BiLSTM + CRF tagger.
   1. BiLSTM:
   2. CRF:

#### Experiment

##### Textual IE

The authors conducted experiments on two NER datasets: the CoNLL-2003 dataset (CONLL03), and the CHEMDNER dataset for chemical entity extraction. We follow the standard split of each corpora. A word level graph was created with two types of edges:

1. Nearby edges: forward and in reverse edges are made between adjoining words in each sentence, permitting neighborhood logical data to be used.
2. Non-nearby edges: re-events of a similar token other than stop words are associated, with the goal that data can be spread through, empowering worldwide consistency of labeling.

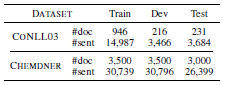


Figure 4 - Results

##### Social Media IE

We build two datasets, EDUCATION and JOB, from the Twitter corpus. The first corpus contains a great many tweets produced by ≈ 10 thousand clients, where the schooling and occupation makes reference to are clarified utilizing far off oversight. We test the tweets from every client, keeping up the proportion among positive and negative posts. The got EDUCATION dataset comprises of 443,476 tweets created by 7,208 clients, and the JOB dataset contains 176,043 tweets produced by 1,772 clients.

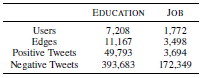


Figure 5 - Statistics for EDUCATION and JOB datasets

A chart is built as self-image organizations, i.e., when data around one client is extricated, a subgraph is shaped by the client and his/her immediate neighbors. Every hub compares to a Twitter client, who is addressed by the arrangement of posted tweets. Edges are characterized by the followed by interface, under the supposition that associated clients are bound to come from a similar college or organization.

#### Results

GraphIE outflanks the SeqIE gauge in many credits, and accomplishes 1:2% improvement in the miniature normal F1 score. It affirms that the advantages of utilizing format diagram structure in visual data extraction.

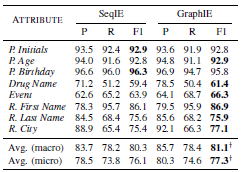


Figure 6 – Results

### Information extraction from text messages using data mining techniques

The point of this paper is to gauge the self-destructive inclinations of an individual by applying information mining procedures to the instant messages an individual ships off the related individuals. By breaking down the segments of the instant messages (watchwords and emojis) the self-destructive propensities of an individual can be assessed so essential advances can be taken to save the existence of the subject.

#### Dataset

The data is procured by isolating every one of the texts send by the subject. This can be refined from different sources like Facebook, WhatsApp, etc. Every one of the messages send through these educating organizations are taken care of in a data base where the model can be applied and assessment for the inclinations ought to be conceivable. The instructive assortment will contain text sort of data and emoticons. No other kind of data, for instance, pictures will be examined through the model.

#### Model

##### Sentiment Analysis

In this part the information is appointed an assessment, for example, positive or negative and the degree of it by performing information pre-preparing utilizing SVM calculation. Text pre-preparing is needed because of the idea of the dataset. The model comprises altogether of the accompanying segments:

* Tokenization
* Data Standardization
* Emoji Conversion
  + Positive Emoji
  + Negative Emoji
* Stop-word Removal
* Stemming
* N-gram
* Term Frequency
* Inverse Document Frequency
* SVM
* K-NN

#### Results

The outcome got from the proposed model gives the assessed slant expectation of the subject dependent on the instant messages sent by the client. The subsequent yield can be utilized as a rule, the psychological issues and anxiety is assessed and, in this manner, if there should be an occurrence of "basic" slants the companions and relatives of the subject can make moves to support, persuade and inspire the enthusiastic height of the subject consequently bringing about the amicability and genuine feelings of serenity of the subject. Subsequently, such estimation examination models are a prerequisite for molding the general public into an occurrence place.

## Experimentation

### Dataset

The dataset used will be the Twitter News Dataset available [here](https://users.dcc.uchile.cl/~mquezada/breakingnews/). This dataset contains 5234 news events gained from Twitter. The report events.csv.gz contains a CSV record, called events.csv with all the news events got from Twitter since August, 2013 until June, 2014. The arrangement of each line of the archive is the going with: <event ID>, <date>, <total keywords>, <total tweets>, <keywords> Where:

* <event ID> is a number which recognizes the relating event. There are 5234 events, by then <event ID> goes from 1 to 5234.
* <date> is the date of the event or related section. The arrangement is YYYY-MM-DD.
* <total keywords> is an entire number showing the quantity of watchwords is in the event or related fragment.
* <total tweets> is an entire number showing the quantity of tweets has a spot with this event.
* <keywords> is a string containing <total keywords> expressions. There is a semicolon between two watchwords.

The record tweets.csv.gz (available upon request through email to the makers) contains a CSV archive, called tweets.csv, with all of the tweet’s IDs identifying with each event in events.csv. The association of each line of the archive is the going with: <tweet ID>, <event ID> Where:

* <tweet ID> is a long number showing the Twitter ID of the given tweet. Using the Twitter REST API, it is practical to recuperate all the information about the given tweet.
* <event ID> thinks about to the event ID of the given tweet.

The records cluster\_labels.txt and time\_resolutions.txt contain the gathering marks for each event and the time objectives acquired from all events, independently. cluster\_labels.txt contains one entire number for each line, from 0 to 19. In line I, the gathering name in that line identifies with the event ID number (I). time\_resolutions.txt contains one skimming point number for each line, showing the time objective learned for all events, in minutes. There are 20 numbers in the report, one for each line, in extending demand, with at most 13 decimal numbers after the point.

### Machine Learning

Observing the dataset, it can be seen that not all the features that are available are useful to this research. The dataset however provides keywords in the <keywords> feature of the dataset extracted from the tweets but not the tweets itself. These keywords can be used to design an ML model to help identify and extract major tweets from twitter. Since there are keywords available, an approach that can take advantage of this is used i.e., Naïve Bayes algorithm ().

Let, A be the event that signifies the availability of a major topic;

B be the event that signifies the availability of keywords in a tweet

Naïve bayes algorithm will predict the probability of event A provided that event B has already taken place. This will help us identify based on high probability the major topics from twitter. Although this approach is finalized due to the analysis of the dataset, further research might result in the changing of the dataset to accommodate a classification approach for this research.

# Project Evaluation

The examination of this evaluation will be finished utilizing the going with data:

* The results of the Model will be measured with performance metrics like confusion matrices. The results from the confusion matrix will be compared with results from the literature review of existing approaches.

The performance metrics are subject to change depending on the algorithm/approach used to tackle this research.

# Conclusion

Social media is gaining momentum regarding its popularity for information sharing. This has resulted in social media becoming a hub for sharing news about various topics around the world efficiently. To make use of this huge amount of data for current affairs, approaches to identify major topics from social media are being done. This research aims to identify and extract tweet with major event content using Naïve bayes theorem. The algorithm will be provided with a set of keywords, the availability of these keywords in a tweet will decide the influence of the tweet in the news domain. The machine learning algorithm will be decided after more research on this topic although Naïve Bayes algorithm is a good contender.

# References

1. Martinez, L., Tsou, M. and Spitzberg, B., 2019. A Case Study in Belief Surveillance, Sentiment Analysis, and Identification of Informational Targets for E-Cigarettes Interventions. Proceedings of the 10th International Conference on Social Media and Society.
2. Bontcheva, K. and Derczynski, L., 2016. Extracting Information from Social Media with GATE. Working with Text, pp.133-158.
3. Brambilla, M., Ceri, S., Della Valle, E., Volonterio, R. and Acero Salazar, F., 2017. Extracting Emerging Knowledge from Social Media. Proceedings of the 26th International Conference on World Wide Web.
4. Qian, Y., Santus, E., Jin, Z., Guo, J. and Barzilay, R., 2019. GraphIE: A Graph-Based Framework for Information Extraction. Proceedings of the 2019 Conference of the North.
5. Ahmad, S. and Varma, R., 2018. Information extraction from text messages using data mining techniques. Malaya Journal of Matematik, S (1), pp.26-29.