Hot Topics in My Area

Aman Thakur, B.Tech. and Lipsa Mishra, B.Tech.

A Capstone submitted to University College Dublin in part fulfilment of the requirements of the degree of M.Sc. in Business Analytics

Michael Smurfit Graduate School of Business, University College Dublin

September 2018

Supervisors: Assoc. Prof. Peter Keenan, UCD,
Mr. Eoghan Keegan and Mr. Emmet Hutchin, AIB

Head of School: Professor Anthony Brabazon

Table of Contents

List of Figures

List of Tables

List of Algorithms

Preface

This practicum is done in partial fulfilment of the MSc Business Analytics in the Michael Smurfit Graduate Business School, University College Dublin (UCD). The work has been undertaken in conjunction with Allied Irish Bank (AIB). AIB is one of the supposed "Big Four" business banks in Ireland. It offers a full scope of individual and corporate banking services. AIB wants to understand the usefulness of detecting hot topics in the local area for its marketing department.

This project assesses various topic models to extract a list of terms related to the emerging topics in 7 geographic locations and ranks for each location, the terms by topic, by order of dominance of frequency on the twitter. This project report also provides an aggregate sentiment analysis on the dataset of tweets collected, discusses some recent and classical topic modelling studies and evaluates the approaches used for this study.

Dublin
September 2018

Aman Thakur
Lipsa Mishra

Acknowledgements

First and foremost, we would like to express our sincere gratitude to our project advisor Prof. Peter Keenan for the continuous support of our study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped us in all the time of research and writing of this thesis. Besides our advisor, we would like to thank our Program Director, Prof. James McDermott, for his encouragement, insightful comments, and hard questions.

Our sincere thanks also go to our mentors in AIB, Eoghan Keegan and Emmett Hutchin for their proactive suggestions at every stage of our project, leading us to work on diverse ideas. Additionally, we would like to express our gratitude to Jason Roche from AIB for his last-minute but very useful recommendations on improvising our results and tips on the writing of this thesis.

Last but not the least, we would like to thank our family for supporting us in our every endeavour.

Executive Summary

Social media is a great source of information for businesses and organizations to understand their customers better and in discovering new trending topics. Many social media platforms offer an open developer access to their APIs. In this study, the focus was laid on topic modelling upon the twitter dataset which offers geolocational tagged tweets and tries to discover trending hot-topics for a given location. LDA and CTM were adopted as our modelling techniques and for conducting experiments on various datasets of tweets from distinct locations across the globe. The study is divided into three phases essentially. In the first phase, we focus on gathering data using a twitter developer account and its API, to create datasets of 5000 tweets on 7 different geographic locations for our samples. The second phase is initialized by pre-processing on textual datasets using NLP techniques and later deploying LDA and CTM algorithms along with optimally tuned hyperparameters to extract topics and term distribution matrices. We further move on to an analysis of hot-topics from raw output by using time series analysis plots to identify trending topics and terms. This is also followed by a sentiment analysis against each topic, which is run using the NRC lexicon.

In the third and the final phase of this study, we develop informative visualizations to enhance the understanding of these locations, based topics through time series plots, word clouds, frequency bar plots, LDA inter-topic clusters and tweet maps. To evaluate our results, we have relied on log-likelihood, perplexity and semantic-coherence as the metrics of comparison in tandem with feedback based on human interpretations of the output topics as well.

List of important abbreviations

API	Application Program Interface
CMC	Co-training based Multi-view Clustering
CTM	Correlated Topic Modelling
DCV	Direct Concept Vector
DTM	Document Term Matrix
HTV	Hot Term Vector
KV	Kind of Vector
LDA	Latent Dirichlet Allocation
MCMC	Markov Chain Monte Carlo
MVTD	Multi-Hour Virtual Topology Design
NEV	Named Entity Vector
NLP	Natural Language Processing
PAM	Pachinko Allocation Model
PLSA	Probabilistic Latent Semantic Analysis
PV	Part of Vector
SMC	Stage-based Muti-view Clustering
STVSM	Suffix Tree Vector Space Model
TDT	Topic Detection and Tracking
TF*PDF	TF-Proportional Document Frequency
VSM	Vector Space Model

Chapter 1 - Introduction

1.1 Background

In recent years, social media has been a useful means for users around the world to communicate with each other for their professional and personal interests. With the rise of social media users, it has also become an important source of data for analytics in different disciplines.

1.1.1 Social Media Analytics

There is an enormous amount of information in social media data. Historically, companies paid market research companies to conduct voting and survey groups among their consumers to get the insights that is now freely available on social media. This has led to the rise of a new generation of market research - the mining of unstructured information from social media and the application of advanced machine learning and machine learning algorithms on them to quantify the data for use in businesses. Companies can use this research to gauge customer's perception of their brand, the current trends in the market for their products and any foreseeable market developments in the future. Hence, Social Media Analytics is the collection and analysis of social media data to gain insights to support decision-making in businesses (Rouse, 2017). Broadly, four types of social media analytics have been identified:

- Descriptive analytics:
  Descriptive analytics gives a general idea about the trends and mood of the market. It works by summarizing data like user comments by visualizations, reports and clustering.

- Predictive analytics:
  Predictive analytics forecasts the emergence and future of trends while also focusing on why and how a certain trend will occur.

- Diagnostic analytics:
  Diagnostic analytics uses historical data to retrieve the reasons as to why a certain outcome went wrong or right in the past. It aids the company in knowing if they are taking the right marketing approach.

- Prescriptive analytics:
  Prescriptive analytics is helpful in suggesting the best approach to reach an optimum level. This is particularly useful in scenarios where the analysis is made on a diverse and unpredictable group of social media users.

1.1.2 Impact

Stieglitz et. al (2018) identified that with the help of these metrics, social media can benefit three domains of trades – namely, in businesses, disaster management and political and other journalism.

Businesses can benefit from the detection of new trends, issues that could relate to bad publicity and to maintain a brand persona which could provide an online channel for direct communication with its consumers. Data from analysis could aid in measuring Key Performance Indicators by giving real-time insights for the corporate.

Disaster management is another such field where social media data has a major impact. Social media has been a useful channel for updating the public in and about an affected area with useful information on how to tackle the situation. In the context of analytics, unstructured data shared by the public in these areas can be used to analyse real-time statistics of a disaster and can be used to detect the location of affected individuals as demonstrated in many pieces of research. This has proven to increase the efficiency of emergency management agencies. In other cases, an approach to monitor the spread of a disease by sentiment analysis has also been demonstrated.

Lastly, social media has been an established platform in the recent years for communication in politics and journalism. It has been a channel for the public to debate on issues and their consequences and for political parties and governments to communicate with the users, thereby reaching a broader audience and gaining more followers. Analytics in this domain has been used to detect the public opinion about governments and to formulate strategies for elections using these insights.

1.1.3 Twitter for Analytics

Twitter is a microblogging social networking service where users can interact with short messages called tweets. These messages were originally limited to 140 characters per tweet but since November 2017, the allowance was doubled to 280 characters. Registered users can post tweets whereas everyone who obtain access to the website can read any tweet of their interest.

Evan Williams, one of the founders of Twitter.com, defined twitter as an information network that “delivers to people the best and freshest most relevant information possible. It tells people what they care about as it is happening in the world”. Additionally, Twitter further emphasized their information network strategy by changing the status updates question from” What are you doing?” to” What’s happening?”. Regardless of whether Twitter could be a suitable alternative to an authoritative source of news, it still provides an outlet to real-time data that predates the best newspapers in informing about the emerging topics with many self-proclaimed authors (users) with an impressive response time (Cataldi et. al, 2010).

Although Twitter gained a massive popularity after its launch, its user growth is no longer accelerating. Twitter’s focused approach makes it a great tool for online marketing and it still caters to a wide range of audience. It holds a special interest for researchers as all activity data (tweets, followers etc.) of any individual user is public, it allows a developer to download up to 20000 random tweets in an hour and provides full access to its tweets to large corporates for analysis purposes. Nevertheless, even the rate-limited access to the tweets base has statistical significance in the academia, as a sufficiently large random subset of a substantial dataset will act like its parent set with a high level of measurable certainty.

Twitter’s inherent openness for public consumption, very well documented API, availability of various developer tools makes it an appealing platform for social web mining. The micro-blogging site is especially intriguing for this purpose since tweets are published at the "speed of thought" and is accessible for utilization as they occur. They speak to the broadest cross-area of society at a global level and are intrinsically multifaceted. Tweets and Twitter's "following" mechanism interface individuals in an assortment of routes, going from short conversational exchanges to interest graphs that associate individuals and the things that they think about. Twitter's asymmetric model makes it stand apart from other social sites like Facebook and LinkedIn which require the mutual acknowledgement of an association between users whereas Twitter enables one to stay aware of the most recent happenings of any other user (Russell, 2013).

1.2 Statement of the Problem

The problem under the study of the capstone is that of detecting trending topics from a geographical area. Trending topics, in this case, are a collection of words that occur across tweets from similar themes. The words in question would then help to identify the emerging trends or any trending event in the geographic area. Further by the identification of a trend, insights into a certain trend was expected.

Businesses can capitalize on the identification of any emerging trends in their domain. For example, the latest trend in fashion could benefit the retail business in their production strategy or an emerging health scare could similarly apprise the health industry to better plan their stock management for the required supplies.

1.3 Purpose of Study

The purpose of this study is to investigate and establish the usefulness and accuracy of a topic modelling algorithm on a corpus mined from twitter and it's the ability to predict trends or events in a specific geographical location.

1.3.1 Study Objectives

The objective of this study is twofold:

1. To stream tweets filtered based on a specific geographic location and identify the trending topics.
2. To recommend and demonstrate a suitable use-case of our model for the benefit of any business.

1.4 Research Question

The research question addressed in this study is:
"Can topic modelling and sentiment analysis on geotagged tweets be used to identify hot-topics in an area?"

1.5 Scope of the Study

There are many popular online social networks and likewise, there was a huge choice for obtaining the data for the research. For this project, data was obtained for the social media site of Twitter as it is easy to access geotagged data for use for research from a Twitter development account. Twitter has a 3-tier model for the use of its APIs, out of which 2 are paid and provide access to deeper analytics. However, for the project, an unpaid and standard developer account was used. The standard Twitter API available to a developer has limits in the number of calls an account can make for the access of tweets. Theoretically, a user can access a maximum of up to 3200 tweets per API call via an account. Further, data collected was restricted to a few specific geographic locations for a comparative analysis of the insights. Lastly for the purposes of our experiment, modelling was done with a maximum limitation of 5000 tweets per location.

1.6 Limitations of the Study

Limitations in this study include:

1. The large volume of disparate data makes it difficult to discover relevant topics, trends and events and in some cases, it is hard to make a definite conclusion unless the topic in question is very dominant over the location. Twitter's specific character limit of 280 words further make this constraint significant as the tweets contain very little information to gain reliable insights in case of absence of any dominant trend over the tweet corpus.

2. Studies report consistently that 1– 2% of tweets include geo-tagging information. This makes it harder for the study to obtain unbiased and relevant

results for the experiment – it runs the risk of not being the voice of the public. (Stieglitz et al., 2014).

3. It is challenging to quantify the performance of a social media research as the topics obtained must be identified based on the user's inference as the algorithm does not have the human intuition to identify the inherent theme of the results obtained. Furthermore, it is difficult to judge the success of such a model as the topics generated are dynamic and change with the changing data in the tweets over periods of time.

1.7 Structure of Project

Chapter 2 is the business background that the project is aimed at and introduces the idea behind using the model studied, for marketing purposes by different businesses upon successful implementation.

Chapter 3 is a literature review and covers an introduction to the research done in social media analytics and the idea of hot topic detection. It also covers the relevant work done in other studies - which have used a similar concept of detecting trends by their algorithms, it discusses the results achieved by them and the evolution of algorithms in this field over the years.

Chapter 4 introduces the methodology used in this study. It includes an overview of the implementation process, a description of the software and algorithms employed, of the data sets collected, and the methods used for preparing it for use by the algorithm. Chapter 5 discusses topic modelling and sentimental analysis whereas Chapter 6 documents the experimental work based on the use of collected data sets and the models reviewed in Chapter 4.

Chapter 7 is an analysis of the results obtained in Chapters 4 and 6 and Chapter 8 covers the conclusions and suggestions for future work.

Chapter 2 - Business Background

Social Media Analytics “listens” to available user-generated social media content, rather than actively asking for user input and acting upon it for supporting business activities. (Holsapple et al., 2014). In case of our project, we are interested specifically in problem/ opportunity detection with the help of our model that may help marketing departments in many businesses.

Nowadays, organizations are cautious with their promotions; by foreseeing customer reaction and keeping away from unforeseen blunders to avoid a viral purchaser backlash in online networking systems. Social media has a twofold role in the field of marketing. It enables organizations to converse with their demographic and, in the meantime, it enables regular users to converse with each other. Moulding clients' exchanges to guarantee they are aligned to the association's objectives is the company's best move for social media marketing. At the same time, disappointed clients can challenge boisterously, turning away numerous different clients effectively and harming the brand's image (Saravanakumar and Suganthalakshmi, 2012).

With the use of a hot topic detector, a company can remain abreast of any anomalies related to their brand image in a specific area – in case of viral backlash and at the same time, get leads on any business opportunities by keeping a track of what the consumers are talking about in an area and discovering effective ways to monetize the current mood of an area.

Chapter 3 - Literature Review

3.1 Social Media Analytics

Online social networking websites have witnessed unprecedented growth in their user base. This has led to a huge change in how individuals share, collaborate, create, and consume data and has made social media a critical information system. From a research perspective, social media can be considered as a live lab, where researchers now have access to the massive quantities of information produced by and about people, things, and their interactions in a real-time environment (Stieglitz et al., 2014).

The power of social media analytics rests less on the fact that the data that is big than it is about the capacity to collect, monitor, analyse, summarize, and visualize the social media data. It is usually driven by specific requirements from a target application. Social media analytics research serves several purposes including extracting useful patterns and intelligence for use by several entities (Zeng et al., 2010).

For instance, sentiment analysis on social media is being often used by politicians now, to check the public feeling on strategies and political positions, to identify churn in previously loyal political subjects, and to manage their persona online. In the event of cataclysmic events, social media can be used to track situational data from affected areas. Additionally, data can be utilized by consumers to make more educated choices while purchasing. Furthermore, social media data can be analysed for intellectual insights on how users in various sections, geographic or social, respond to major worldwide events etc. In recent years, this field of research has developed rapidly and continues to create and assess strategies, specialized algorithms that can mine the crude data from these online sources for conversion into useful information that can suit business purposes (Stieglitz et al., 2014).

3.2 Hot Topic Detection

Topic Detection attempts to endeavours to distinguish "topics" by investigating and arranging the content of textual materials, empowering the aggregation of different snippets of information into manageable clusters automatically. In the context of news, Topic Detection can be viewed as an event detection that gathers stories into a corpus, wherein each such gathering is related to a single topic. Meanwhile, a hot topic is defined as a topic that is dominant in the social media over a stretch of time. The hotness of a topic depends relies up on two elements: how frequently the terms show up and the number of records that contain those terms (Chen and Seng-Cho, 2007).

Chen et. al (2002) demonstrated that no subject can stay hot indefinitely; at the end of the day, every topic is subject to the Aging theory - it goes through a life cycle of birth, development, and demise. Hence, each trending topic evolves over a given period. They showed this relevance in terms of news items, but this is also true for hot topic detection using twitter analysis.

Long et. al (2011) emphasized the importance of summarization after the tracking of hot topics and suggested methods for it. It is easy to display a list of words representing the trending topics during a period. However, it makes little or no meaning if the reader of these topics has no background knowledge of what trends can be inferred from the topical words. Therefore, it is necessary to provide a comprehensive summary along with the hot topics.

With their study, Long et. al (2011) introduced the method of event tracking as a through a maximum-weight bipartite graph matching algorithm and proposed generating a summary with several relevant posts while also considering their topical coverage and ability to account for the variability in popularity over time.

3.3 Related Work

Hot topic detection is one of the significant research fields for public sentiment analysis and the key methodologies include text classification and clustering, topic detection and tracking, sentiment analysis and the automatic summarization of multiple documents. Below is a brief discussion on some of the approaches that have been used for this purpose.

3.3.1 TF-Proportional Document Frequency

Bun and Ishizuka (2002) were one of the earliest to work on an emerging trend detection algorithm for mined online texts. They used online news archive from 4 newswire sources concurrently for two separate weeks as their data source and their objective was to create an automated weekly summary of the main topics covered by the sources.

Their approach was found to be useful in summarizing the main topics for a periodic set of data and for extracting the terms explaining the main topics. It made use of the fact that the prominent topics in the news would also be the most discussed ones on multiple newswire sources. For summarizing the topics, they used sentence vector clustering and for recognizing the terms associated with the significant topics, the TF-Proportional Document Frequency (TF*PDF) algorithm was used (Bun and Ishizuka, 2002).

The TF*PDF algorithm assigned heavy term weight to the most frequently cited terms in the documents. The terms with highest term weight can also be referred to as unit vectors. The unit vectors helped to unveil the words explaining the hot topics for a group of concurrent texts. Each sentence in a text was associated with a sentence vector consisting of the unit vectors. Further, a topic-wise clustering was done with the sentences with high average weight using their sentence vectors. At the end of their experiment, they were able to identify the emerging topics, classify the sentences according to their respective topics and arrange them chronologically to form a summary of the topic with a success rate (Bun and Ishizuka, 2002). However, it failed to consider the variation in popularity of a topic over time.

3.3.2 Probabilistic Latent Semantic Analysis (PLSA)

A widely adopted approach for topic modelling, based on parametric probability distributions is probabilistic latent semantic analysis or probabilistic semantic indexing. The PLSA technique suggests that a document label and a word are conditionally independent for an unobserved topic. (Blei et al., 2003), (Hoffman,1999).

The PLSA technique skips an overly generalizing assumption made by the unigram model, which is that every document shall be made from a single topic. Therefore, it is able to handle the possibility that a document may have multiple topics (Blei et al., 2003).

However, a drawback with PLSA technique is that it is not a well-defined generative model of the given documents and there is no organic way to use it to allocate probabilities to a new document, moreover, the number of parameters which must be estimated grows linearly with the number of training documents (Blei et al., 2003).

3.3.3 Pachinko Allocation Model (PAM)

Many topic models like LDA fail to capture correlations between the topics. Another emerging topic modelling technique by Blei and Lafferty (2006), Pachinko Allocation Model (PAM), brought out these arbitrary, nested and sparse relationships be-tween topics through the help of directed acyclic graphs (DAG). Here, the leaf nodes of a directed acyclic graph would represent individual words in the vocabulary. At the same time, every internal node would represent a correlation among its child nodes which are the related sub-topics (Li and McCallum, 2006).

However, few internal nodes can also have children nodes that may represent different topic or a mixture of them. Through the presence of multiple nodes in a DAG, PAM can capture not just correlations among words, but also the correlations among topics, which LDA could not (Li and McCallum, 2006). For example, in a document with discussion upon the following topics - cooking, health, insurance and drugs. Cooking could seldom occur about the topic of health; however, health, insurance and drugs could be discussed in a similar context. PAM can describe these correlations with “a directed acyclic graph using 4 nodes for the 4 topics, from 1 level that is directly connected to the words” (Li and McCallum, 2006).

3.3.4 Timeline Analysis and Multidimensional Sentence Modelling

Chen et al. (2007) also used a dataset of text-based news for a definite window of time. However, their objective was an improvisation on the TF*PDF because it considered the variation in the ‘hotness’ of a topic over time. They addressed the fact that a topic normally appears at a high frequency, over a specific time period, until it is trending or hot and gradually reduces in frequency as it ages. They introduced two important factors of a hot term in their logic - pervasiveness and topicality, which improved the quality of extracted results.

Pervasiveness accounted for the frequency with which a term appeared in a set of documents and topicality accounted for the variation in this frequency of usage over a period. The weight of a hot term was the summation of these two variables for that term and it took both a high TF*PDF value and a good life cycle variation to create a heavy weighted term. This combination of TF*PDF and the Aging Theory led to a more accurate ‘hotness’ extraction for a term, countering the drawback of a dominant TF, regardless of its hotness. However, identifying keywords is not enough to piece out an emerging topic. Therefore, a further clustering approach needs to be applied. For this purpose, they used five kinds of sentence vectors - HTV, NEV, DCV, PV, and KV for sentence modelling as opposed to one vector for sentence representation. They ran several experiments using this approach and compared it with previously tried approaches. The results indicated that their algorithm led to significantly more genuine hot topic extractions (Chen et al., 2007).

3.3.5 Hot-stream

Phuvipadawat and Murata (2010) also contributed to the field of tracking and topic detection (TDT) by formulating a method to collect, group, rank and track breaking news on Twitter. To improve the clustering results, they increased the scores for proper nouns and each cluster was then ranked on popularity and accuracy. Using this approach, they created an application where users can discover breaking news from the Twitter timeline.

Since this study only focused on #breaking news and narrowed the number of users referred for their tweet base, it was concluded that it might have different results if we adopt this approach for a broader section of detection.

3.3.6 Temporal and Social Terms Evaluation

Cataldi et. al (2010) were one of the earlier researchers to work on twitter data for hot topic mining. Their methodology consisted of extracting and formalizing the multilingual user-generated tweets as vectors of terms and mapped each term with a relative frequency.

They then used the well-known Page Rank algorithm used by Google to evaluate the social relationships in the network and determine the authority of the users. Based on this information, a set of emerging terms was selected and the keywords depending on their life status were defined. Finally, they connected the emerging terms with other semantically related keywords, thereby producing a navigable topic graph. This accommodated the detection of trending topics and worked for user-specified time constraints(Cataldi et. al, 2010).

Their technique was able to form broad clusters of emerging topics and did not consider a user's profile, preferences or specific events. The social relationships in a user network were also studied and the significance of each analysed tweet was quantified. The keyword-based topic graph connected the hot topics with their co-occurrent ones. This allowed for the detection of the trending topics for a certain period, that was user-specified. (Cataldi et. al, 2010)

3.3.7 Frequent Pattern Stream Mining

Twitter had gained massive popularity by 2012. Guo and Guo (2012) experimented with another approach for hot topic detection from twitter data. They attempted to modify a pre-existing method, taking in consideration that the Twitter data was different than other previously mined textual datasets. An individual tweet is short, sparse and rapidly spreading - making it not as easy as, for instance, a news source, for finding the most frequent terms in a data stream of tweets. They proposed a more flexible and practical approach for the mining of frequent topics on Twitter - the Frequent Pattern Stream Mining algorithm and suggested modifications to the basic algorithm of it to make it adaptable to the nature of twitter data.

The basic FP-Stream algorithm can detect time-sensitive and frequent patterns from data streams in restricted main memory. However, it could not be used for this instance of hot topic mining using Twitter streams as FP-Stream processes its data in chunks and only works when there is a suitable amount of data available in one data chunk, each precedent and subsequent data chunk being the same size. This is impractical in a real-world scenario as it would delay the data processing until a given size of data is obtained whereas public opinion analysis is usually dependent on a period, rather than the availability of significant amount of data (Guo and Guo, 2012).

In their study, Guo and Guo (2012) revised the FP-Stream structure and embedded a tilted-time window table. This tilted-time window table accounted for three different factors - the time-point information in a stream, the item-set number and the batch number, both in corresponding time. This extended structure made the FP-Stream algorithm more practical and flexible. It was now able to process twitter streams batch by batch with each batch number subject to change with time. Experiments using this algorithm demonstrated its adequacy for the real-world twitter data mining.

3.3.8 Social Topic Detection and Geographic Clustering

In their study, Kim et al. (2013) used a geographic clustering analysis across the states of US based on social topics. They detected the social hot topic by using the ratio of word frequency. For this analysis, geotagged public statuses tweeted in North America regions were collected and the objective was to evaluate how people shared about the trending social events across different areas in the US. Using this methodology, they were able to identify geographic communities which had a strong correlation to each other by studying the time series for a set of topic words. The visualized result using the Google Fusion Table was able to demonstrate this in more clarity. Along with the demonstration of the usefulness of using a clustering algorithm based on a social topic in classifying social communities, this study was also able to realize the effectiveness of using ratio of word frequency for social events or news detection and suppressed the regular but small talks and sentimental words such as lol, like, and love.

This was one of the earliest examples of using geographically divided twitter data amongst our readings and since our study would be to identify hot topics across a region, it was found to be of significant interest.

3.3.9 Multiview Clustering

The performance of most topic detection methods is severely restricted due to the short length of a tweet, the use of random and sometimes newly invented abbreviations in their content. Fang et al. (2014) proposed the use of multi-view clustering to address these problems using a Multi-Hour Virtual Topology Design (MVTD) algorithm. This approach collated the multiple relationships that can exist between a group of tweets to detect topics, that is it considered not only text semantic relations, but also semantic relations, social tag relations and temporal relations.

Upon examination on real datasets, it was found that execution of MVTD is vastly improved than that of a single view, and it is valuable for distinguishing themes from Twitter. This approach took the following nature of tweets in consideration - the interesting topics usually consist of consecutive phrases, activities like comment and retweets will be considerably more noteworthy than those for subjects getting less consideration, and hotly debated issues typically show up with numerous hashtags and much of the time, the related tweets of a specific theme show up in a specific scope of time and districts, with a couple of themes accepting consideration for quite a while from individuals everywhere throughout the world (Fang et al., 2014).

In this study, a new document similarity measure based on a suffix tree (STVSM) was used to measure the relations of each pair of tweets. It exploited the extracted meaning carried by phrases. They enhance the tag feature space by regarding words appearing in tags as tags too and employed three multi-view- based methods, which are word-tag, SMC and CMC to do the clustering. Their experiment demonstrated the superior performance of multi-view clustering in comparison to any single-view clustering and led to better hot topics detection from Twitter (Fang et al., 2014).

3.3.10 Combined Content and Time Similarity

Amongst the other hot topic identification methods that we came across our research was one based on combined similarity. Zhao et al. (2017) combined several methods of calculating similarities like cosine similarity and Jacobi formula of the content. They also calculated the similarities both in context of content and time to get the final similarity. Their approach was based on three different models - linear model, quadratic polynomial model and neural networks. After the similarity calculation, clustering was implemented using an incremental algorithm called Single-Pass algorithm. The speed of clustering was fast, and the topic detection was highly accurate using this method. However, this method was found to be only suitable for large-scale document collection as it required enough similarity threshold to achieve reasonable results.

3.3.11 Latent Dirichlet Allocation (LDA)

Huang et. Al (2012) proposed using single-pass clustering method with Latent Dirichlet Allocation (LDA) instead of Vector Space Model (VSM) model which is traditionally used for emerging topics extraction. They used a Markov Chain Monte Carlo (MCMC) to generate the samples from the dataset of tweets and then LDA was applied to this generated sample of tweets to find the texts like each other. LDA is a probabilistic model that considers the document as a 'bag of words' and ignores the order in which the words are written. The LDA model was found to be more accurate for calculating text similarity over VSM, reducing the rate of hit and miss alarm rate. Hence it was found to be more effective in reducing the effect of data scarcity which the VSM faces and made the topic search more focused.

The general thought of Latent Dirichlet Allocation (LDA) depends on the theory that a man composing a report has certain themes in his mind. To expound on a point for a certain theme means picking a word with a specific likelihood from the pool of expressions for that theme. An entire record can then be identified as a blend of various subjects. At the point when the creator of a report is one individual, these subjects mirror the individual's perspective of a record and her specific vocabulary. When different users are talking on the same topic, the subsequent themes mirror a collaborative shared perspective and the labels of the subjects mirror a typical vocabulary to depict the record. LDA clarifies the likeness of information by gathering highlights of this information into imperceptible sets (Krestel et al., 2009).

Chapter 4 - Methodology

4.1 Process Overview

The scope of this project lies in determining the hot topics in each location through analysis of the tweets that Twitter users have been sharing within the premises of the given location. Once the emerging topics are extracted from tweets, we also determine the sentiment of the tweets against their identified topic. In our study, the range of locations considered can be as wide as an area within a city, such as an area like Blackrock in Dublin or as large as a city such as London, UK.

The overall methodology can be divided into 3 phases:

- Phase 1: Aggregating Data (creation of tweet bank)
- Phase 2: Data Processing (hot topic extraction)
- Phase 3: Data Visualization (visualize topics and sentiments)

4.2 System Overview

The system developed using the R programming language was composed of different components. The Dataset Aggregation module utilizes twitter's API for R to query desirable tweets based on location given by user input and saves the dataset in a compressed data file in the local disk. The Pre-processing Module then reads the compressed file and brings it in memory as a data frame.

The data frame is passed through a data cleaning pipeline which involves text normalization and preparation tasks and further creates a document-term matrix from it. The LDA module and CTM module create the desired topic model using the given optimal parameter settings, including the iteration count, burning parameter, keep parameter, number of topics and number of terms per topic.

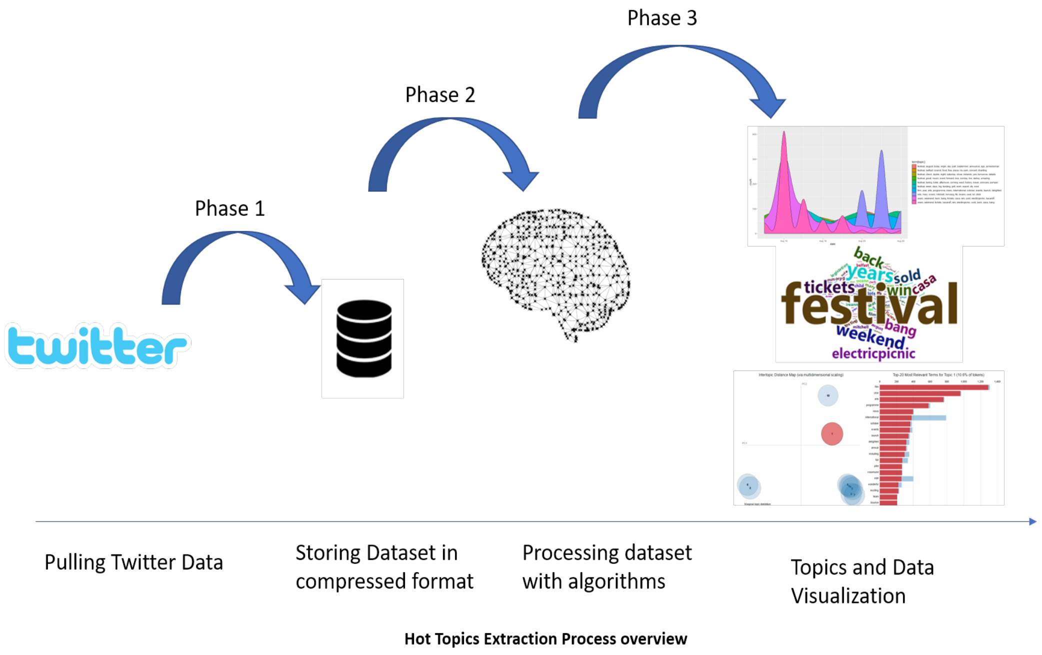

Phase 1
Phase 2
Phase 3
twitter
back
years
sold
tickets
win
casa
festival
bang
weekend
electricpicnic
Pulling Twitter Data
Storing Dataset in compressed format
Processing dataset with algorithms
Topics and Data Visualization
Hot Topics Extraction Process overview

Figure 1: Overall Architecture from Tweet to Topics

In the first phase, we utilize the existing APIs available to us through Twitter. We set up a free developer account to obtain the necessary API keys. Using this API, we query the Twitter's platform to fetch tweets and its related metadata, amongst which, the location metadata is relevant for the experiment. There are restrictions associated with each call that is made using the Twitter API and the number of such calls that can be made from one Twitter developer account is finite. Thus, aggregation of tweets has been a gradual process, whereby the scripts in R were run multiple times to fetch the tweets and accumulate geolocation-based tweet datasets.

4.3 Why these Techniques were Adopted

In the Data Processing phase, we deploy and test 2 major topic modelling algorithms, namely, Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM), by tuning them in different control parameter and settings. The topic models deployed are based on statistical distribution - the Dirichlet and the Logistic Normal distribution. In our scope, a latent topic discovery approach has been used and hence the underlying mechanism of topic modelling has been clustering based unsupervised machine learning methods. The classification based supervised learning algorithms have not been deployed as it does not fit our research objective. Since tweets are highly unstructured documents and small in length, with a maximum length of 280 characters, a classification based supervised learning model does not suit our purpose.

4.4 Data Collection

We rely heavily on the existing APIs provided by Twitter for creating a pipeline to fetch tweets and metadata on regular frequent intervals. We use 'twitteR', an R based open source library. The search function in the above libraries helps us query the platform with specific geographical latitude and longitude and a desired number of tweets to be retrieved in a JSON format, which later is converted to a data frame for initial storage.

4.5 Dataset Quality

In our experiments specifically, the quality of the dataset is determined by the key parameters in question, namely, tweet text and its geolocation. However, there are other parameters such as date range in a search query and quality of the tweet text (less Non-ASCII characters). It was observed that most of the tweets fetched from Twitter's API are not geolocation tagged by their users. The average percentage of tweets with their latitude and longitude tagged is only between 10 % to 15% for each of the used datasets. Although Twitter also approximates the geotag of its tweet based on the IP address and account location, it is not as precise and reliable.

4.6 Dataset Description

Each dataset that is used in our experiments has been gathered in a range of 10-30 days and has tweets from a location such as a metropolitan city or a larger area such as a state. Each such dataset initially had several attributes for each tweet. A few of them have been described as below:

- Created at – mentioning the time of creation of tweet
- Id – an integer representation of the user id
- Id_str – the string representation of the user id
- Text – actual text in the tweet
- Source – the medium used for posting the text, for example, a cell phone

- in_reply_to_status_id – if the tweet is a reply, this contains the representation of the original tweet
- in_reply_to_user_id - if the tweet is a reply, this contains the representation of the original tweet's author
- User – author of the tweet in question
- Coordinates – Longitude and Latitude of the user if geotagging is turned on for them, or an approximated geographic location by Twitter
- Place – if the tweet is related to a place
- quoted_status_id – if the tweet is a quote
- retweeted_status - if the tweet has been broadcasted by other users

However, for the purposes of our modelling, we only used the relevant characteristics – the text, the time of creation of tweet and the coordinates for each of the tweet.

4.7 Data Storage

The data frame file is compressed into an '.RDS' file, which is in a hyperdata format for easy access and storage in the local disk.

4.8 Data Preparation with Natural Language Processing (NLP)

In the data preparation step, we intend to first extract the raw text out from the JSON structured text and then clean it according to our pre-processing requirements. In this step, tokenization is introduced where we break down the structure into the relevant constituent tokens, which are, individual word level and sentence level tokenization. We need not go deeper for a paragraph level tokenization as Twitter has a limit of 280 characters per tweet.

While data preparation, we intend to apply the standard text mining pipeline to prepare our data for analysis to be done upon. The text mining pipeline is a standard process that ensures the transformation of a document into its resultant document-term matrix. In our case, each tweet represents a document, however, in terms of their properties are different. Documents usually have massive amounts of text and are organized in structural paragraphs. Tweets, however, are restricted to only 280 characters and can be semi-structured. A typical tweet consists of hashtags, usernames, slang terms and URL to other image, web pages or videos. Tweets can also be in regional languages. However, for the current topic modelling problem we have restricted only to English language tweets.

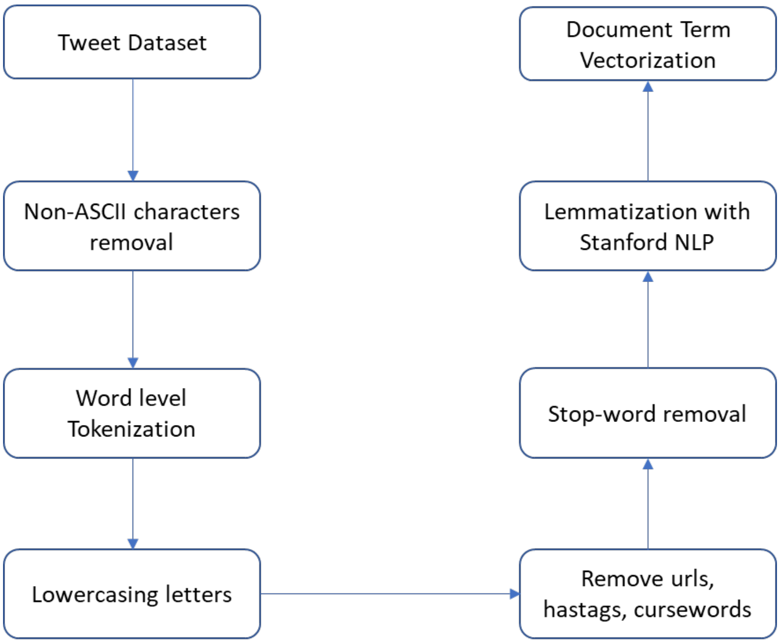

Tweet Dataset
Non-ASCII characters removal
Word level Tokenization
Lowercasing letters
Remove urls, hastags, cursewords
Stop-word removal
Lemmatization with Stanford NLP
Document Term Vectorization

Figure 2: Data Preparation Process

Once the tokenization of the tweet is complete, we get the words, separated by space out of the tweet. We intend to use NLP library sci-kit learn to achieve this step. The next step in the process involves removal of the stop words from the tweet. This helps further in condensing high entropy information by removal of conjunctions such as ‘a’, ‘an’, ‘or’, ‘but’ and other similar connecting words. The tokenization step is followed by lemmatization of the words where each word from the tweet is lemmatized to its root word, for example, (have, having, had) shall lemmatize to have. The lemmatization step helps us to reduce complexity from the text and aid in the filtration at a subsequent stage. In this pipeline, stress must be laid on filtering out words based on their frequency. Thus, as in the following data preparation stage, we put a low-frequency filter on the remaining words in the tweet and remove such words. However, determining the cut-off frequency for this stage is key. A low-frequency filter could end up in entropy loss if a certain keyword is only mentioned once in a tweet. Thus, figuring out the dynamics of this step shall be achieved with more trials and empirical assessment.

4.9 Topic Analysis

In the former stages we have acquired the requisite data, stored, cleaned and prepared it to reach a level from where different algorithms can be tested and deployed on. In the data analysis stage, we aim to reach from a bunch of geotagged tweets to figure out the topic that is categorized into a higher-level topic that has the highest semantic alignment amongst them. We have narrowed to 2 key algorithms for processing to the higher-level topic models, namely, LDA (Latent Dirichlet Allocation) and CTM (Correlated Topic Model).

In the above context, it is important to tune optimal hyperparameters for the 2 algorithms. The alpha value, beta value, number of topics (k) and the iterations are the basic hyperparameters for LDA along with the type of sampling methodology.

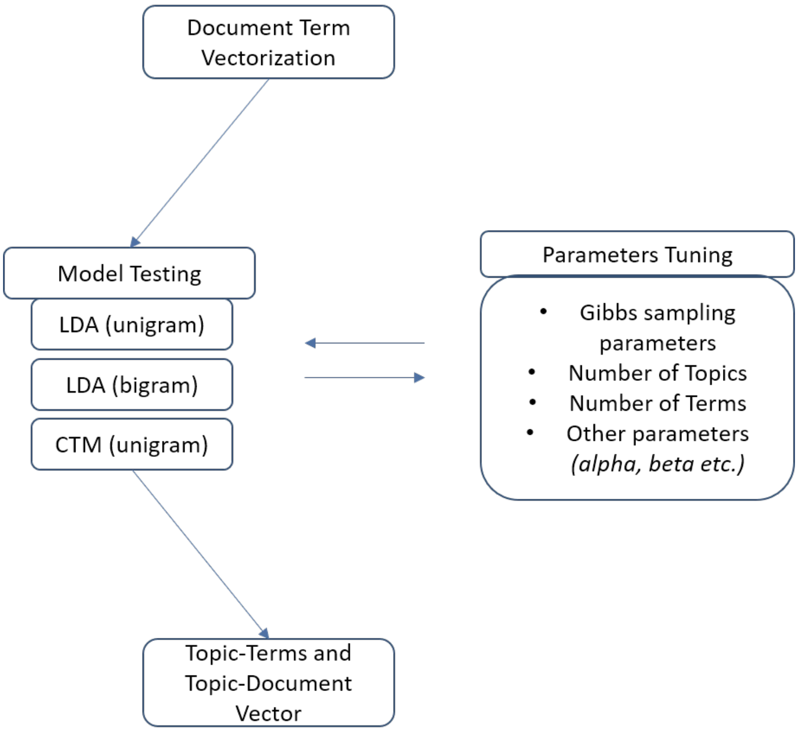

Document Term Vectorization
Model Testing
LDA (unigram)
LDA (bigram)
CTM (unigram)
Parameters Tuning
• Gibbs sampling parameters
• Number of Topics
• Number of Terms
• Other parameters *(alpha, beta etc.)*
Topic-Terms and Topic-Document Vector

Figure 3: Data Processing and Topic Modelling Approaches

We have already discussed how we progress from raw tweet text to applying, cleaning and tokenization techniques over it. It is followed by 2 other steps – lemmatization and normalization, where we aim to condense the information, perform stop words removal and aim to streamline the remaining words into lowercase.

4.10 Software Tools

To develop the computer program for fetching geolocational tweets, process them with different models and visualize them, we have used the R programming language and its open source libraries. The development environment, runtime environment and code editor used for the program was R Studio.

4.11 Open Source Libraries

The following are the key open source libraries used in the R program.

- 'twitteR' – for querying and fetching geotagged tweet data
- 'ggmap' – for plotting maps
- 'ggplot2' – for plotting various graphs
- 'tm' – for all kinds of text-processing
- 'topicmodels' – for building topic models
- 'ldaVis' – for inter-topic cluster maps
- 'writexl' – for writing program output to spreadsheet

Chapter 5 - Topic Modelling and Sentiment Analysis

5.1 Topic Modelling

A topic model helps to discover hidden topical pattern across a document. It is based on a statistical generative model. This model considers the word co-occurrence relations among documents and uses this to classify the documents into several soft clusters called topics. It has been applied on knowledge discovering and latent semantic analysis very successfully in the past decade. (Wen et. al)

Topic models assume that the content of documents is governed by hidden and abstract topics. In most cases, topic models assume that each document is a bag of words. It is also assumed that the order of words is unimportant, that each document is assigned several topics and that documents are otherwise independent (Lozano et al., 2017).

5.1.1 Latent Dirichlet Allocation (LDA)

LDA has been a common method for identifying topics from online media for the past decade. It is a generative topic model and assumes that each word in a document is generated from a topic that in turn is picked from the topic distribution for each document. The topic distribution for each document is generated from a Dirichlet distribution which means that Latent Dirichlet Allocation allows a document to consist of several topics of different magnitude (Lettier, 2018).

A simplified explanation of how the LDA algorithm works is as follows. A unique set of words is collected, as corpus and fed to the algorithm. In this scenario, it was the processed document after the tweets were mined. The number of topics, k to be generated is then decided upon. For our model, we used harmonic means to decide the optimum number of topics to be generated. The number of words to be generated per topic was decided on, two hyperparameters of alpha and beta were used, each being a number between zero and infinity (Lettier, 2018).

The alpha hyperparameter controls the mixture of topics for any given document. The lower it is, the less likely for a document to have a vast mixture of topics and vice-versa. Similarly, the beta hyperparameter controls the distribution of words per topic. The lower it is, the smaller number of terms are allocated for a topic. The algorithm then generates a list of terms for every topic. It also builds terms vs topics table and topics vs document probability distribution tables (Lettier, 2018).

5.1.2 Corelated Topic Model (CTM)

LDA assumes each of the topic generated by the model to be nearly independent of each other, which leads to the strong and unrealistic modelling assumption that the presence of one topic is not correlated with the presence of another. This leads to an inability of the model is to find a relationship amongst the topics themselves. Correlated Topic Model was introduced by Blei and Lafferty (2006) to mitigate this issue. In simple terms using CTM, words like astronomy and stars are more likely to appear together in a document rather than in conjunction with a word like cat (Bagdouri, 2011).

This model builds on the earlier latent Dirichlet allocation and does not assume an independence between the topics and instead focuses on the correlation between them. The limitation of LDA arises due to the use of Dirichlet distribution to model variability between topics, whereas CTM uses the more flexible logistic normal distribution to exhibit correlations amongst the topics. This incorporates a covariance structure among the components and gives a more realistic model of the latent topic structure, enabling the construction of topic graphs and document browsers that allow a user to access the documents in a topic-guided manner (Blei and Laffterty, 2006).

The goal of topic modelling is to predict unseen items dependent on a set of observations. To conclude in brief, an LDA model will predict a list of words based on the latent topics that the observations suggest, but CTM can predict items associated with additional topics that are correlated with the conditionally probable topics. The latter model is a hierarchical model of document collections and like LDA models the words of each document from a mixture model, instead of a bag of words approach. Additionally, like LDA, the mixture components (unigrams or bigrams) are shared by all documents in the collection and the mixture proportions are document-specific random variables. CTM allows each document to exhibit multiple topics with different proportions and can thus capture the heterogeneity in grouped data that exhibit multiple latent patterns (Blei and Laffterty, 2006).

5.2 Key Metrics

Even though the latent space found by topic models is significant and helpful, assessing such models is difficult because topic discovery is an unsupervised procedure. There is no highest quality level comparison of topics to look at against for each corpus. Accordingly, assessing the latent space of topic models needs us to accumulate exogenous information (Chang et al., 2009). Below are the explanations for the metrics we used to evaluate the performance of our models.

5.2.1 Log-likelihood

Log-likelihood measures the probability of how well a model fits the observed data. The higher the likelihood, the better the model for the given data. The overall log-likelihood of a model is calculated by taking the harmonic mean of the log likelihoods in the Gibbs sampling iterations after a certain number of “burn-in” iterations. Several studies have raised concerns about the accuracy of this method but note that it reports the ranking of the models correctly, which was useful in terms of quantifying which of our models worked the best (Konrad, 2017).

5.2.2 Perplexity

Perplexity is also a measure of model quality and in topic modelling is often used as “perplexity per number of words”. It describes how well a model predicts a sample. The lower the score, the better the model for the given data. “It can be calculated as exp^(-L/N) where L is the log-likelihood of the model given the sample and N is the number of words in the data” (Konrad, 2017).

5.2.3 Semantic Coherence

Log-likelihood and perplexity do not relate to the context and semantic coherence between words. For a good topic model, it is important to have meaningful terms in the output so that a user can relate them to a topic. Hence, we considered the topic coherence measure for this evaluation. Simply put, each topic consists of words, and the topic coherence is considered for the top N words from the topic. It is defined as "the average/median of the pairwise word-similarity scores of the words in the topic" (Rosner et al., 2014). Topics with high topic coherence scores will belong to a good model, generating coherent topics.

5.3 Sentiment Analysis

Sentimental Analysis is "a process of computationally identifying and categorizing emotions and opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a topic, product, etc. is positive, negative, or neutral" (Gupta, 2018). For our model, we used the Syuzhet package. The package comes with four sentiment dictionaries and we used the NRC lexicon which provides eight different emotion scores and a positive/negative indicator as the result. A detailed explanation of the result from this package is discussed in section 7.1. Sentiment Analysis is the most commonly used modelling technique used for Twitter analytics these days and we combined this on top of topic modelling to get an overall emotional idea of what the hot topics assessed were expressing.

Chapter 6 - Experiment

6.1 Model Testing

The official API provided by Twitter for the R programming language allows making calls to their database through their web service for fetching tweets with desired search parameters. The key parameters used while querying and fetching datasets were:

- 'search term'
- 'number of tweets'
- 'geo-coordinates (latitude, longitude)'
- 'radius size (km)'
- 'since date'
- 'until date'

API calls for requested geocoordinates and in the given date range parameters were very restrictive. On a practical note, most Twitter users do not actively turn on their location services on their computer or smart devices, due to which twitter engine is not able to record their geo-coordinates. To enrich its location-based data collection, twitter also approximates user locations based on their account location (country), historical activity on the platform and IP address-based approximations.

6.2 Dataset Aggregation

The models were tested on different sizes of locations. In the table below is a description of a few metropolitan cities from all corners of the world that have been used to create the city level datasets. The specifics of each of these city-level datasets are summarized in the table below.

ID	Location	Radius (km)	Geocoordinates	Corpus Size	Dates Range
1	Dublin, Ireland	80	53.349804, -6.260310	5000 tweets	July-August 2018
2	London, United Kingdom	100	51.507351, -0.127758	5000 tweets	July-August 2018
3	New Delhi, India	100	28.613939, 77.209023	5000 tweets	15 August 2018
4	Kerala, India	200	10.850516, 76.271080	5000 tweets	August 2018
5	New York, USA	150	40.712776, -74.005974	5000 tweets	July-August 2018
6	San Francisco, USA	150	37.780079, -122.420174	5000 tweets	July-August 2018
7	Louth, Ireland	80	53.9508, -6.5403	5000 tweets	July-August 2018

Table 1: Details of Dataset used

6.3 Topic Models Deployed

6.3.1 LDA Model

To test this model, the datasets listed in the previous section were used. Latent Dirichlet Allocation model with unigrams builds upon a document-term matrix which is generated using unigram tokens only. To create the document term matrix for test purposes, initial data cleaning and pre-processing have been achieved on the tweets in multiple steps to achieve the suitable tweet corpus. Natural Language Processing Techniques deployed in the pre-processing stage include tokenization, stop-word removal, non-ASCII character removal, lemmatization and low-frequency term filtering.

The control parameters used for tuning the LDA unigrams model for desirable performance are summarized in the table below.

Method	Burning Value	Iterations Value	‘Keep’ Parameter	Number of Topics	Alpha Value	Beta Value
Gibbs Sampling	1000	1000	50	50	0.1	0.05

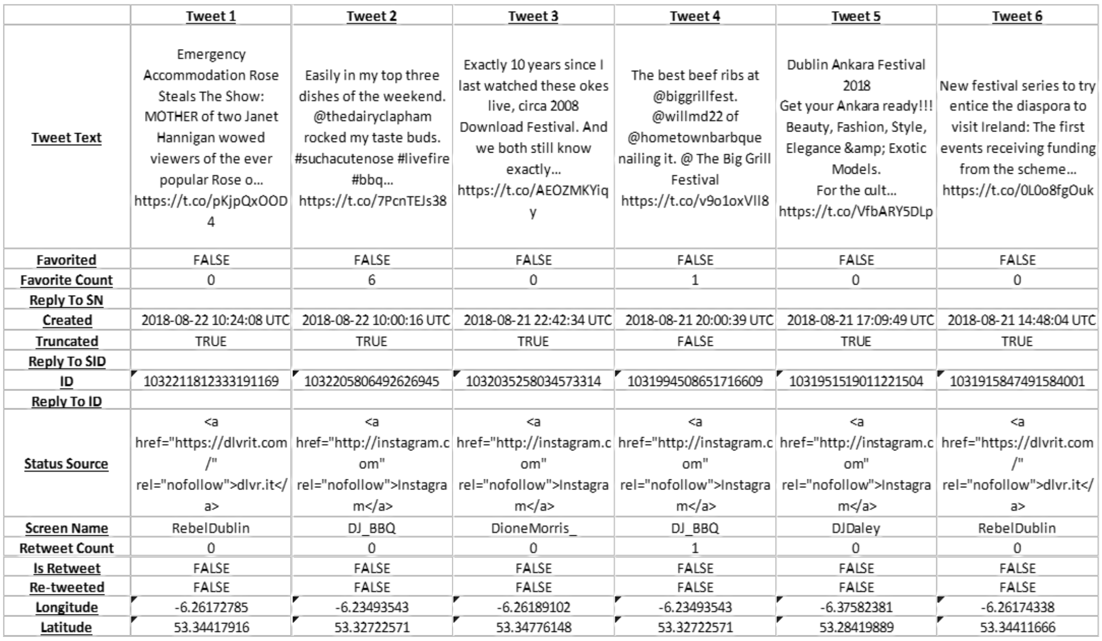
Table 2: Control parameters for LDA Unigram

6.3.2 CTM Model

The Correlated Topic Model is different than LDA in its assumption that it considers the degree of correlation that can occur within k different topics. In order to test the CTM on tweet datasets, a similar methodology has been followed for initial data cleaning and text pre-processing as described in section 6.3.1 for the LDA Model.

6.3.3 Structure of Raw Input Dataset

Tweets datasets are loaded one at a time into the program from local disk and uncompressed back to an R data-frame object. The structure of raw input, with the key important attributes such as ‘Tweet text’, ‘Longitude’ and ‘Latitude’ is depicted in the table below. It is to be noted that the number of tweets considered for testing the models is limited to 5000.



	Tweet 1	Tweet 2	Tweet 3	Tweet 4	Tweet 5	Tweet 6
Tweet Text	Emergency Accommodation Rose Steals The Show: MOTHER of two Janet Hannigan wowed viewers of the ever popular Rose o... https://t.co/pKjpQxOOD4	Easily in my top three dishes of the weekend. @thedairyclapham rocked my taste buds. #suchacutenose #livefire #bbq... https://t.co/7PcnTEJs38	Exactly 10 years since I last watched these okes live, circa 2008 Download Festival. And we both still know exactly... https://t.co/AEOZMKYiqy	The best beef ribs at @biggrillfest. @willmd22 of @hometownbarbque nailing it. @ The Big Grill Festival https://t.co/v9o1oxVIl8	Dublin Ankara Festival 2018 Get your Ankara ready!!! Beauty, Fashion, Style, Elegance &amp; Exotic Models. For the cult... https://t.co/VfbARY5DLp	New festival series to try entice the diaspora to visit Ireland: The first events receiving funding from the scheme... https://t.co/0L0o8fgOuk
Favorited	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Favorite Count	0	6	0	1	0	0
Reply To SN						
Created	2018-08-22 10:24:08 UTC	2018-08-22 10:00:16 UTC	2018-08-21 22:42:34 UTC	2018-08-21 20:00:39 UTC	2018-08-21 17:09:49 UTC	2018-08-21 14:48:04 UTC
Truncated	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE
Reply To SID						
ID	1032211812333191169	1032205806492626945	1032035258034573314	1031994508651716609	1031951519011221504	1031915847491584001
Reply To ID						
Status Source	<a href="https://dlvrit.com/" rel="nofollow">dlvr.it</a>	<a href="http://instagram.com" rel="nofollow">Instagram</a>	<a href="http://instagram.com" rel="nofollow">Instagram</a>	<a href="http://instagram.com" rel="nofollow">Instagram</a>	<a href="http://instagram.com" rel="nofollow">Instagram</a>	<a href="https://dlvrit.com/" rel="nofollow">dlvr.it</a>
Screen Name	RebelDublin	DJ_BBQ	DioneMorris_	DJ_BBQ	DJDaley	RebelDublin
Retweet Count	0	0	0	1	0	0
Is Retweet	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Re-tweeted	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Longitude	-6.26172785	-6.23493543	-6.26189102	-6.23493543	-6.37582381	-6.26174338
Latitude	53.34417916	53.32722571	53.34776148	53.32722571	53.28419889	53.34411666

Figure 4: Sample of abbreviated raw data

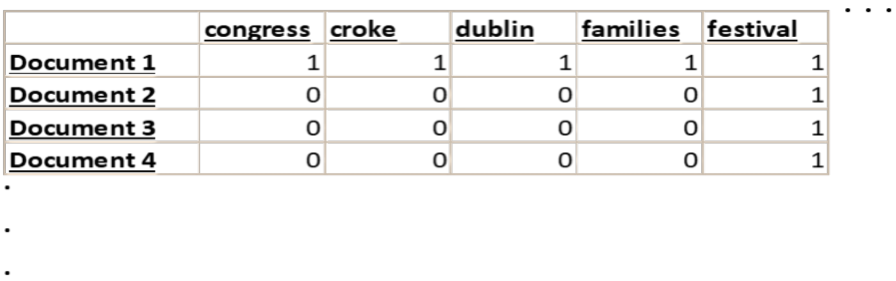
6.3.4 Structure of Processed Dataset

The table depicted in section 6.3.4 is processed using pre-processing techniques such as tokenization, stop-word removal, non-ASCII character removal, lemmatization and low-frequency term filtering as discussed in chapter 4. To vectorize our data and prepare it as an input for the LDA and CTM topic model algorithms, we create a document-term matrix with the characteristics as described in the table below.

Size	**Non-sparse entries**	**Sparsity**	**Weighting Type**	**Dataset ID**
Documents = 5000, terms = 8336	35760/41644240	99%	term-frequency	Dataset 7

Table 3: Description of Document Term Matrix

Below is a screengrab that depicts the DTM for dataset 7. The document -term matrix has a size of 5000x8336 elements for the dataset 7. Each tweet is referred to through its unique id in the rows of the DTM and each unigram/bigram is referred to in the columns.



	congress	croke	dublin	families	festival
Document 1	1	1	1	1	1
Document 2	0	0	0	0	1
Document 3	0	0	0	0	1
Document 4	0	0	0	0	1

Figure 5: DTM for dataset 7

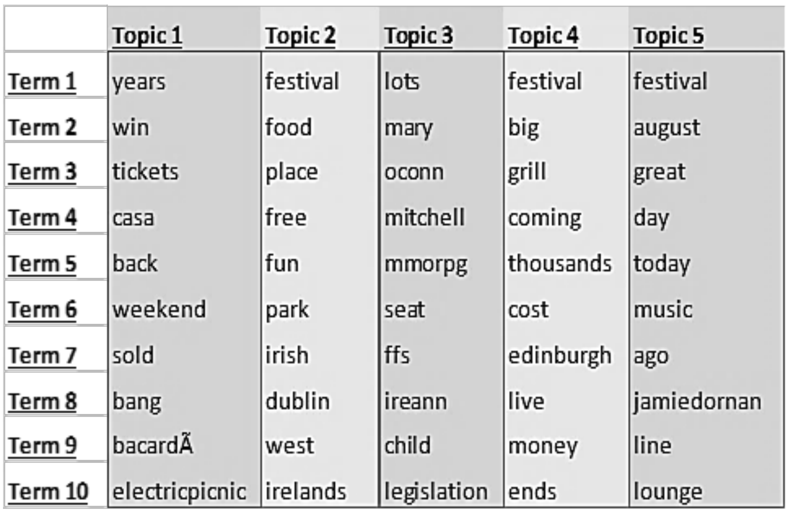
The DTM is also processed before being fed to the LDA or CTM algorithm by identifying any document values in which the column sum is zero. Such documents are removed from the DTM.

Chapter 7 - Analysis

In this section, the output generated by the program for topic modelling and sentiment analysis is presented in the relevant tables, charts and other visualizations.

7.1 Structure of Model Output

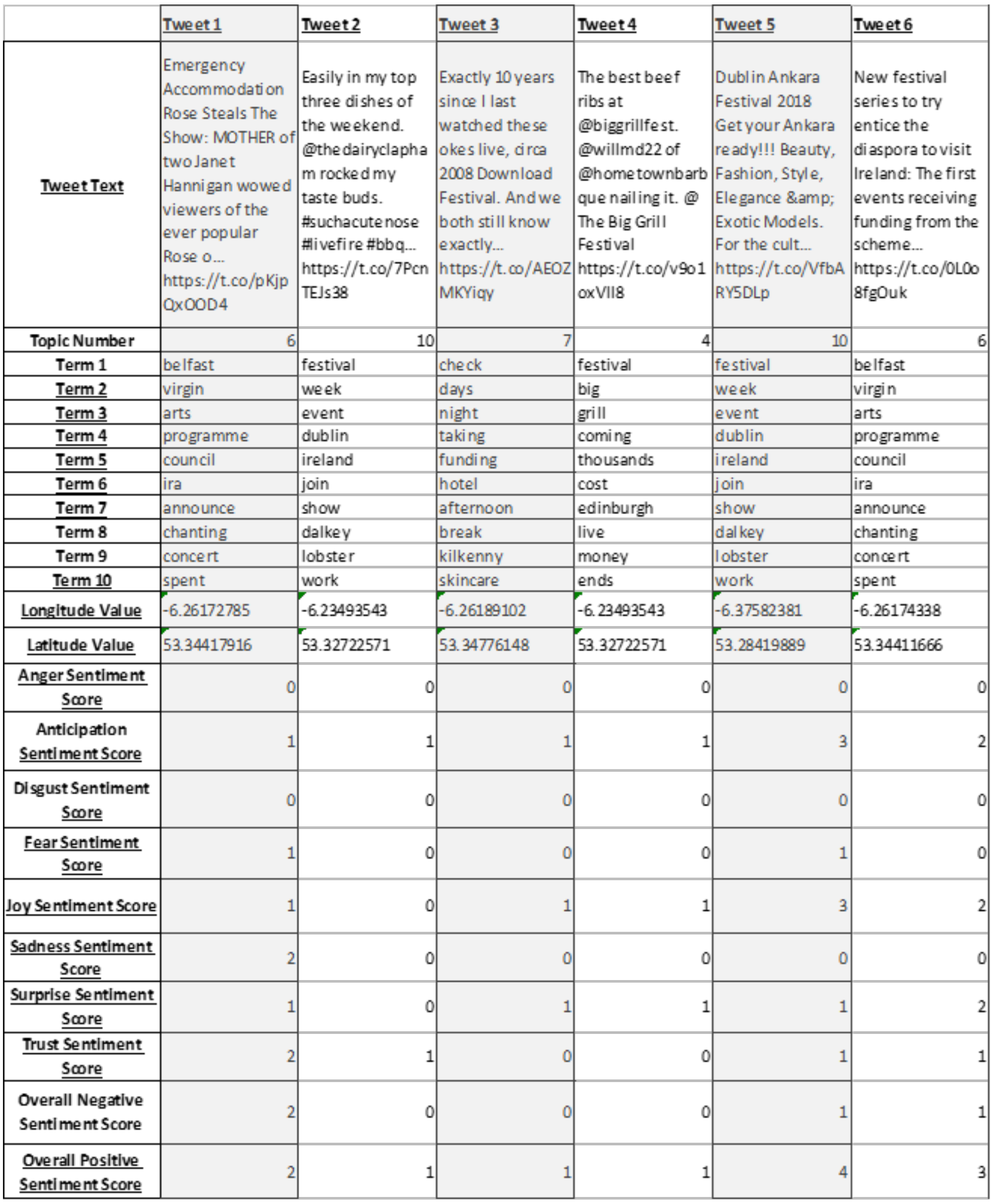
The LDA and CTM models create a document-topic matrix and a topic-term matrix as an output for the given dataset 7 and with the tuned values of hyperparameters listed in the section 6.3.1. Below is a screengrab of the output structure of the topic-term matrix as generated by the LDA model generated for producing 5 topics (k=5) and 10 terms per topic (t=10).



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Term 1	years	festival	lots	festival	festival
Term 2	win	food	mary	big	august
Term 3	tickets	place	oconn	grill	great
Term 4	casa	free	mitchell	coming	day
Term 5	back	fun	mmorpg	thousands	today
Term 6	weekend	park	seat	cost	music
Term 7	sold	irish	ffs	edinburgh	ago
Term 8	bang	dublin	ireann	live	jamiedornan
Term 9	bacardÃ	west	child	money	line
Term 10	electricpicnic	irelands	legislation	ends	lounge

Figure 6: Output structure of topic-term matrix

Further, as an output for a given dataset, the program is also capable of producing the desired topic model and sentiment scores against such topics. The output is written in the .xlsx format to a spreadsheet in a structure described in the table below.



	Tweet 1	Tweet 2	Tweet 3	Tweet 4	Tweet 5	Tweet 6
Tweet Text	Emergency Accommodation Rose Steals The Show: MOTHER of two Janet Hannigan wowed viewers of the ever popular Rose o… https://t.co/pKjpQxOOD4	Easily in my top three dishes of the weekend. @thedairyclapham rocked my taste buds. #suchacutenose #livefire #bbq… https://t.co/7PcnTEJs38	Exactly 10 years since I last watched these okes live, circa 2008 Download Festival. And we both still know exactly… https://t.co/AEOZMKYiqy	The best beef ribs at @biggrillfest. @willmd22 of @hometownbarbque nailing it. @ The Big Grill Festival https://t.co/v9o1oxVII8	Dublin Ankara Festival 2018 Get your Ankara ready!!! Beauty, Fashion, Style, Elegance &amp; Exotic Models. For the cult… https://t.co/VfbARY5DLp	New festival series to try entice the diaspora to visit Ireland: The first events receiving funding from the scheme… https://t.co/0L0o8fgOuk
Topic Number	6	10	7	4	10	6
Term 1	belfast	festival	check	festival	festival	belfast
Term 2	virgin	week	days	big	week	virgin
Term 3	arts	event	night	grill	event	arts
Term 4	programme	dublin	taking	coming	dublin	programme
Term 5	council	ireland	funding	thousands	ireland	council
Term 6	ira	join	hotel	cost	join	ira
Term 7	announce	show	afternoon	edinburgh	show	announce
Term 8	chanting	dalkey	break	live	dalkey	chanting
Term 9	concert	lobster	kilkenny	money	lobster	concert
Term 10	spent	work	skincare	ends	work	spent
Longitude Value	-6.26172785	-6.23493543	-6.26189102	-6.23493543	-6.37582381	-6.26174338
Latitude Value	53.34417916	53.32722571	53.34776148	53.32722571	53.28419889	53.34411666
Anger Sentiment Score	0	0	0	0	0	0
Anticipation Sentiment Score	1	1	1	1	3	2
Disgust Sentiment Score	0	0	0	0	0	0
Fear Sentiment Score	1	0	0	0	1	0
Joy Sentiment Score	1	0	1	1	3	2
Sadness Sentiment Score	2	0	0	0	0	0
Surprise Sentiment Score	1	0	1	1	1	2
Trust Sentiment Score	2	1	0	0	1	1
Overall Negative Sentiment Score	2	0	0	0	1	1
Overall Positive Sentiment Score	2	1	1	1	4	3

Figure 7: Output of topic model in spreadsheet

The documents, which are the tweets in our case are arranged as columns and their requisite output attributes are arranged in the corresponding rows. Each tweet has a corresponding topic number which indicates the topic having the highest probability assigned to it. After the topic numbers, the top t terms (t=10, in this case) are displayed for the assigned topic. For example, tweet 1 for the given dataset 7 has been assigned to topic number 6 as given in the output spreadsheet. The top 10 topics assigned to topic 6 are {'belfast', 'virgin', 'arts', 'program', 'council', 'ira', 'announce', 'chanting', 'concert', 'spent'}. Further to this, the sentiment scores are listed for respective topics. Each sentiment score has a value ranging from 0 to 10 and is calculated using the NRC (National Research Council, Canada) Lexicon for words in the English language.

For the given tweet 1, the overall positive and overall negative sentiment scores are 2 and respectively. The NRC lexicon draws associations for thousands of words in 8 categorical emotions and scores each of them. For tweet 1, the ‘Anger Sentiment Score’ = 0, ‘Anticipation Sentiment Score’ = 1, ‘Disgust Sentiment Score’ = 0, ‘Fear Sentiment Score’ = 1, ‘Joy Sentiment Score’ = 1, ‘Sadness Sentiment Score’ = 2, ‘Trust Sentiment Score’ = 2, as per the NRC Lexicon. The R program also writes the associated ‘Longitude Value’ and the ‘Latitude Value’ for each tweet, to the output spreadsheet, which for the tweet 1 are Longitude = ‘- 6.26172785’ and Latitude = ’53.34417916’.

7.2 Data Visualizations

7.2.1 Tweet Map

Data visualization are integral to visually summarize the trends in the data through charts, maps and other visuals. The R program outputs the respective geocoordinates of the tweets in the output spreadsheet.

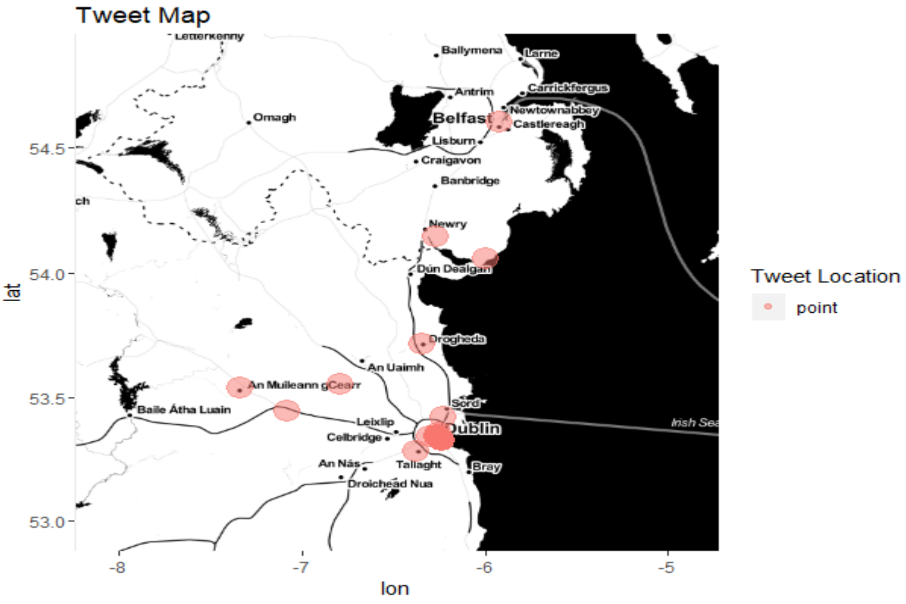


Figure 8: Tweet map

Using the google map's open source R library (ggmap) in tandem with ggplot2, the above map is created to visualize the locations of the individual tweets. The red coloured points in the above map mark individual tweet location and their density in a region is depicted by the strength of the coloured of the representative point.

7.2.2 Time Series Graph for Topics

Time series analysis helps track the change in the trend of a variable over time and helps keep a track of the fluctuations. Using the R program, a user can output a time series graph with 'topic count' on the y-axis and the 'date range' on the x-axis as depicted in the figure below.

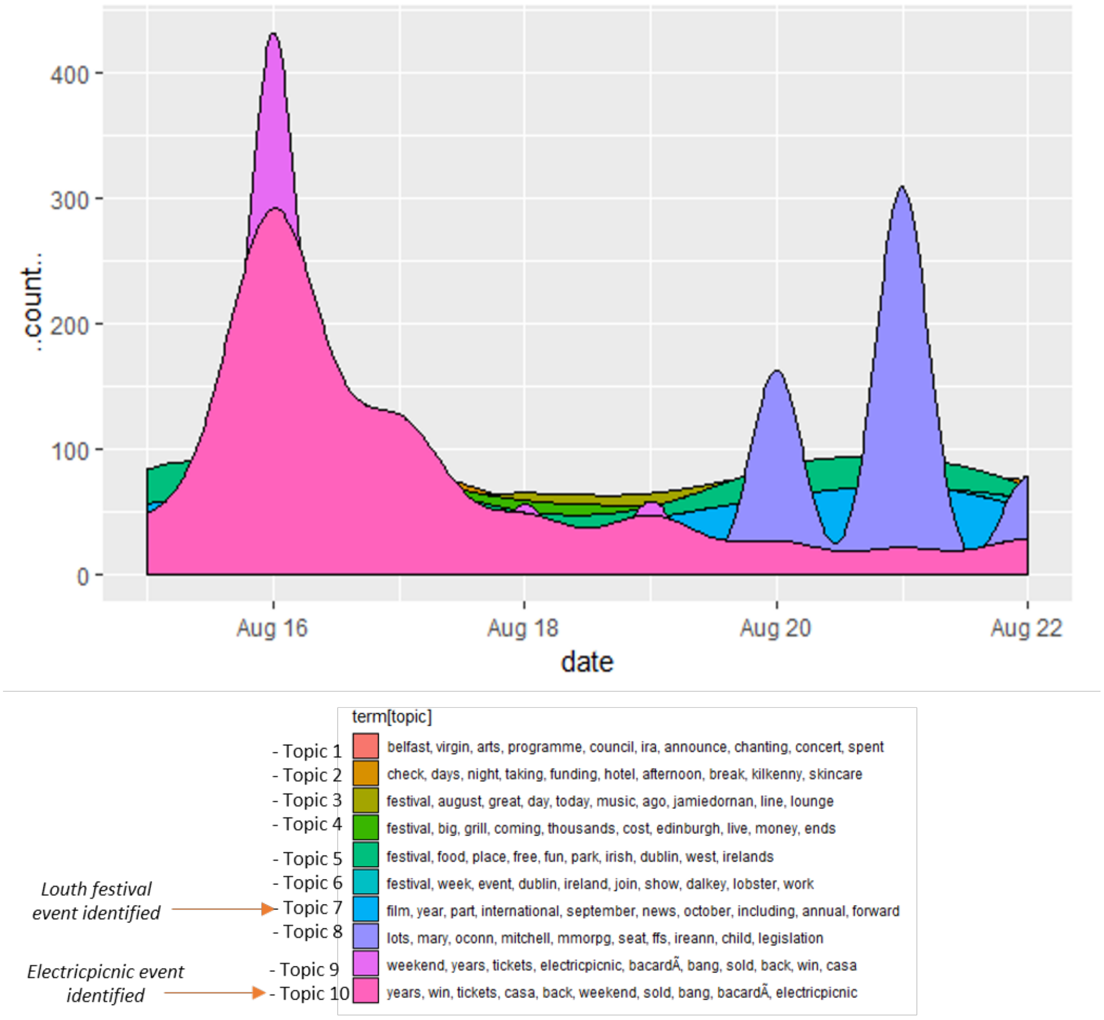


Figure 9: Time-series model of trending topics

The time series graph is helpful for the users to investigate the traction that different topics have gathered amongst Twitter users for a queried location in the past few days. To identify the difference between a topic and a 'hot-topic' for a location, the above graph can be helpful. Topics which have consistently high counts of discussion over a short period of time are co-relatable with events which are either happening in an area or are about to occur.

7.2.3 High Frequency Word Cloud and Distribution

In contrast to topic models, the high-frequency term visualizations aren't as reliable for investigating the overall tweet dataset, however, they do visually represent the important words that compose the tweet corpus and ongoing discussions in a location. The program hence produces a bar plot of high-frequency words and a word cloud to visualize the same.

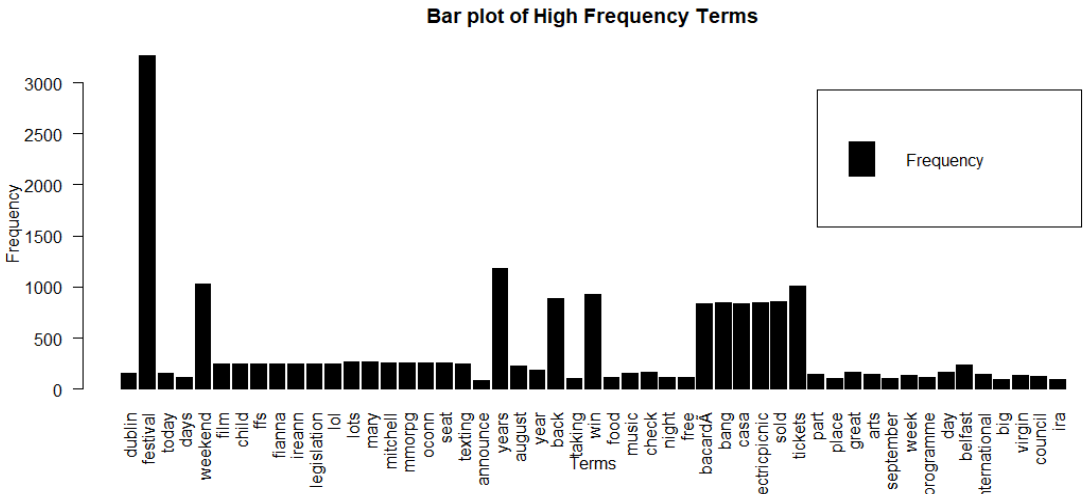


Figure 10: High frequency word distribution

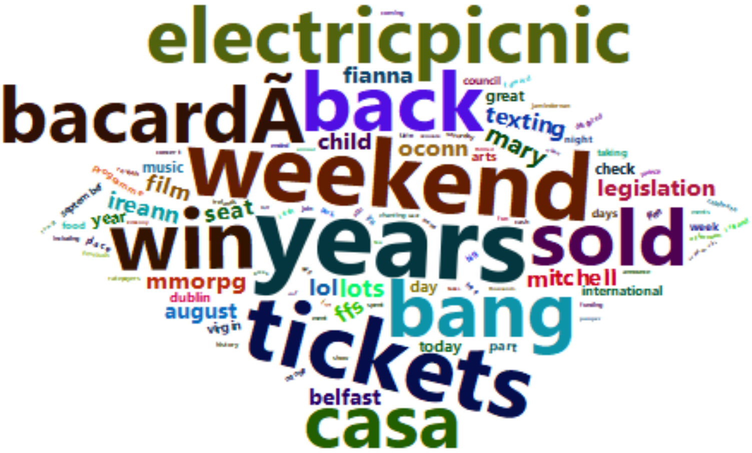
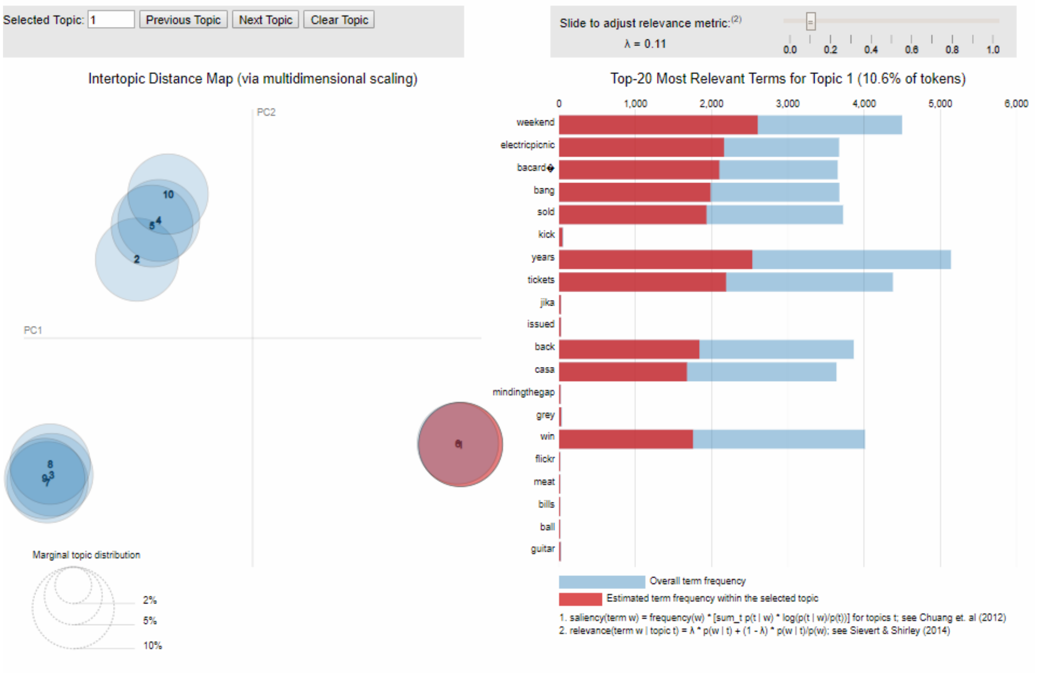


Figure 11: High frequency word cloud

The word cloud in the above figure is built for the dataset 7 and the size of each word is scaled to its frequency of occurrence in the given tweet corpus. The ggplot2 and wordcloud2 libraries in R are used for generating the same and a user can hover over a word to see its frequency count in the program.

7.2.4 Inter-topic Cluster Map

Both LDA and CTM are generative probabilistic models and hence their output topics can be visualized on component axis as clusters using appropriate visualization tool. The program uses the 'ldaVis' library to create an inter-topic cluster visualization for a given 'k' (k =10, in this example) important topics in the tweet dataset. The visualization is adjusted to depict 't' top terms for each of the 'k' topics.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 12: Interactive visualization for K topics

It is to be noted that in the above visualization,

- The sizes of the topic circles are proportional to their relative count in the tweet dataset.
- Clicking over a term on the right sub-part of the figure will highlight all the topics that comprise of the selected term
- Lambda (λ) is a step size for inter-topic distance.
- In the left sub-part of the figure, the overall frequency of each term is plotted in blue colour while the estimated term frequency within the topic is plotted in red colour.

7.2.5 Sentiment Analysis and Scoring

The latter part of our analysis deals with the identification of the sentiment which is associated with each topic and tweet in the dataset for a given location. The 'syuzhet' library in R has been used for sentiment analysis and score calculation for every tweet and topic. The library uses the NRC (National Research Council, Canada) sentiment lexicon for calculating the sentiments in 8 identified emotion categories,

- 'anger'
- 'anticipation'
- 'disgust'
- 'fear'
- 'joy'

- ‘sadness’
- ‘trust’

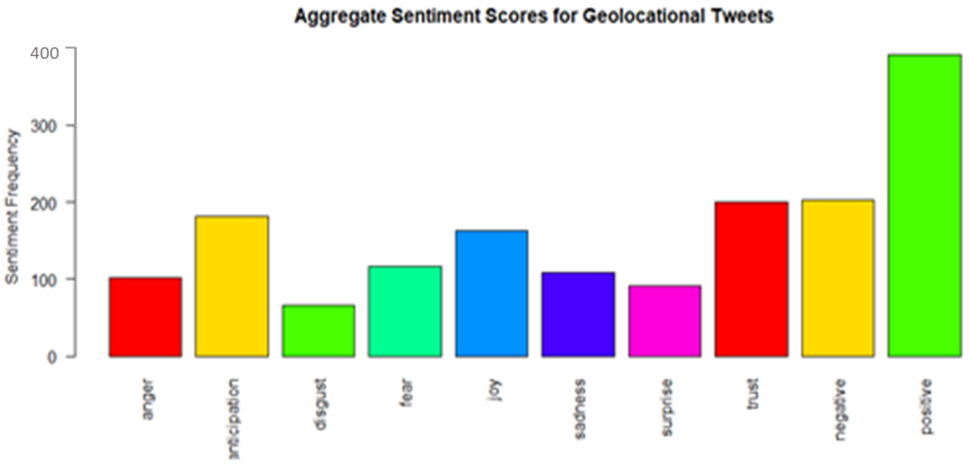


Figure 13: Aggregate Sentiment score for Geo-locational tweets

To analyse the aggregated sentiment score of an entire tweet dataset of a location, the program plots a bar plot as shown above in figure 13. Individual topic and tweet associated sentiment scores are written in the output spreadsheet as depicted in section 7.1.

Chapter 8 - Results and Discussion

We have guided the above experiment in search of the best performing topic model to fit our purpose of extracting inherent semantic topics from the gathered tweet datasets. The LDA model has been fast and has produced desirable results overall, as seen in section 7.2. Moreover, the time-series analysis done on top of LDA topics are relatable to real-world events or ‘hot-topics’ as seen in figure 9 of section 7.2.2. In comparison, LDA with bigrams model seems to perform decently too on the given datasets, however, its runtime is of higher computational complexity and requires a longer time to execute. Moreover, it is difficult to estimate the relevance of bigram terms in each topic as bigrams may not necessarily produce meaningful topics for every dataset and their relevance can vary depending upon the human interpretation of them.

On the other hand, the Correlated Topic Model (CTM) is different in its underlying assumption of drawing on correlation amongst different latent topics. However, for our concerned tweet datasets, CTM does not produce highly desirable results as conversational topics on tweets may not be inherently correlated to each other. This results in the CTM producing distinct topics with many common terms. Such a topic model is biased with an assumption of correlation and hence may overlook many other latent terms with low correlation values that could have been constituted into topics.

Model	Log Likelihood	Perplexity	Mean Semantic Coherence (k=10)
LDA with Unigram	-2.27 x10^5	3.486 x10^3	0.54
LDA with Bigram	-7.90 x10^5	3.975 x10^3	0.54
CTM	- 8.58 x10^5	6.548 x10^3	0.52

Table 4: Comparison of the models

In the above table 4, we summarize some of the key metrics used to evaluate the topic models in question. The ‘Log Likelihood’ is a measure used to see the performance of the model on unseen documents. The higher the value of the ‘log-likelihood’ parameter, the better is the model’s capacity to allocate topics on a test dataset. The second parameter is perplexity, which is a decreasing function of the ‘log-likelihood’ with respect to the count of tokens. Thus, lower values of perplexity are desirable in a model for comparative analysis.

However, these mathematical parameters are incapable of truly evaluating the interpretations that humans can draw out of topic models. Chang et al. (2009) have described in their paper, how the human interpretation of quality of a topic model and the perplexity metrics can vary significantly and even have an anti-correlation with respect to each other. Based on the human judgement of limited evaluators in this experiment, the LDA unigram model seems to be the most indicative in the identification of an ongoing event.

Chapter 9 - Conclusion

9.1 Summary

In this study, we experimented with geotagged textual data from the Twitter platform with the aim to extract emerging trends and hot-topics for a given geographical location. We relied on the Twitter API to fetch our datasets and dwelled deep into exploring the use of Natural Language Processing (NLP) to pre-process our datasets. Further, we deployed the Latent Dirichlet Algorithm (LDA) and Correlated Topic Model (CTM) to extract the latent topics with their terms. An important part of the experiment has been the interpretability of the output topics in terms of their semantic meaning and human interpretations. We have used various data visualization methods such as word-clouds, frequency plots, time series analysis, tweet maps and inter-topic cluster maps to make the results more interpretable and visually appealing.

The results of the experiments on the dataset 4 and dataset 7 were close to our expectations. The LDA unigram model had performed better than the other models and had the lowest perplexity and decent value of mean semantic coherence relative to other models. The NRC lexicon which was used for the sentiment analysis was able to pull the positive and negative sentiment scores from the tweets and added as a coupled information to the topic model.

9.2 Contribution

9.2.1 Academic Contribution

The algorithms used in our experiments are well known for the topic modelling problems and have been used in various other studies as referenced by us in chapter 3. Related studies have also been conducted on twitter data using these and other techniques. However, in this study, we have experimented on a coupled model of topic modelling and sentiment analysis together to prototype a Decision Support System (DSS) which is geo-aware. Some of the used techniques performed well on our datasets, while some did not perform up to our expectations. Thus, we feel that this study adds its small bit to the previously done work in the research area of Twitter-based social media analytics and applicability of topic modelling and sentiment analysis coupled together on the twitter data of specific locations.

9.2.2 Business Contribution

For over a decade, businesses and organizations have been actively leveraging social media analytics in various ways to understand their customers and members in a better way. The Decision Support System prototyped in this study can help in understanding the ongoing trends in each location of interest and help in supporting marketing strategies and operations indirectly. The topic-terms obtained as output can also improve the Search Engine Optimization practices of a business by suggesting geo-aware keywords.

On a different paradigm, another use of our topic modelling and hot topic analysis could be the detection of ongoing trends and topics being discussed on twitter in real-time for a given location during a Disaster Management situation. This may indirectly aid in understanding the issues of the disaster affected population to develop better evacuation plans.

9.3 Significance

While numerous studies have previously been undertaken on a similar research topic, we felt the most significant finding of our research was the discovery of subtopics. In the case of dataset 4, the location in question, Kerala was suffering from floods during the period of our data collection. After the hot topic analysis was done, we not only detected the trending topics as sentiments expressed on Twitter about the flood – which was expected, but also there were other subset conversations in addition. For example, a few trending topics were related to the applause of the rescue efforts conducted by government and the local fishermen. Other discussions pertained to praising the overseas funds collected by non-residents for the rescue effort. When coupled with sentiment analysis on these tweets, the general emotion of a location can be analyzed, and a business or humanitarian organization can easily assess the different directions in which a population can think and process such or similar events.

9.4 Future Work

The models and dataset used in the study fulfilled our objective and we could build a proof of concept of the ‘hot-topic’ extraction decision support system. One of the major constraints we faced in terms of gathering datasets was the limited availability of geolocational tagged tweets. It was felt that a premium twitter developer account could give a wider access to the Twitter data through unrestricted API calls and help in gathering a larger dataset with a higher number of samples, i.e. geolocational tagged tweets. This might have resulted better than the current study.

Another scope of improvement lies in using a classifier mechanism coupled with the LDA and CTM model. This mechanism can be used to simplify the topic terms of raw output of LDA and CTM and then translate them into broader English terms using an unsupervised classification approach. This shall help in increasing the interpretability of the output topic model and increase its semantic coherence, thereby making it easier for a human interpretation.

Since the interpretation of topics often relied on how well the user was aware of the events in the location, an additional summarization tool can be implemented, using the methods as suggested by Long et al. (2011). This will lead to greater interpretability of the topic terms and better visualizations. For understanding the correlations between hot-topics and news, the available open news APIs of target locations can be integrated with the output of our topic models. This shall be able to pinpoint the likeliness of relation between the evolved twitter hot-topics and existential news article, to give a better interpretation to the sub-topics going on around in a location.

References

Bagdouri, M., 2011. Topic Modeling as an Analysis Tool to Understand the Impact of the Iraq War on the Iraqi Blogosphere.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

Blei, D. and Lafferty, J., 2006. Correlated topic models. Advances in neural information processing systems, 18, p.147.

Bun, K.K. and Ishizuka, M., 2002, December. Topic extraction from news archive using TF* PDF algorithm. In Web Information Systems Engineering, 2002. WISE 2002. Proceedings of the Third International Conference on (pp. 73-82). IEEE.

Cataldi, M., Di Caro, L. and Schifanella, C., 2010, July. Emerging topic detection on twitter based on temporal and social terms evaluation. In Proceedings of the tenth international workshop on multimedia data mining (p. 4). ACM.

Chang, J., Graber, J., Gerrish, S., Wang, C., Blei, D., 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In Proceedings of Neural Information Processing Systems, 2009 (pp. 1-9).

Chen, C.C., Chen, Y.T., Sun, Y. and Chen, M.C., 2003, September. Life cycle modeling of news events using aging theory. In European Conference on Machine Learning (pp. 47-59). Springer, Berlin, Heidelberg.

Chen, K.Y., Luesukprasert, L. and Seng-cho, T.C., 2007. Hot topic extraction based on timeline analysis and multidimensional sentence modeling. IEEE transactions on knowledge and data engineering, 19(8).

Fan, W. and Gordon, M.D., 2014. The power of social media analytics. Communications of the ACM, 57(6), pp.74-81.

Fang, Y., Zhang, H., Ye, Y. and Li, X., 2014. Detecting hot topics from Twitter: A multiview approach. Journal of Information Science, 40(5), pp.578-593.

Guo, J., Zhang, P. and Guo, L., 2012. Mining hot topics from Twitter streams. Procedia Computer Science, 9, pp.2008-2011.

Gupta, S., 2018, January. Sentiment Analysis: Concept, Analysis and Applications. Available at: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications

He, J., Hu, Z., Berg-Kirkpatrick, T., Huang, Y. and Xing, E.P., 2017, August. Efficient correlated topic modeling with topic embedding. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 225-233). ACM.

Hofmann, T., 2017, August. Probabilistic latent semantic indexing. In ACM SIGIR Forum (Vol. 51, No. 2, pp. 211-218). ACM.

Holsapple, C., Hsiao, S.H. and Pakath, R., 2014. Business social media analytics: definition, benefits, and challenges.

Huang, B., Yang, Y., Mahmood, A. and Wang, H., 2012, August. Microblog topic detection based on LDA model and single-pass clustering. In International Conference on Rough Sets and Current Trends in Computing (pp. 166-171). Springer, Berlin, Heidelberg.

Jónsson, E. and Stolee, J., An Evaluation of Topic Modelling Techniques for Twitter.

Kim, H.G., Lee, S. and Kyeong, S., 2013, August. Discovering hot topics using Twitter streaming data social topic detection and geographic clustering. In Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on (pp. 1215-1220). IEEE.

Konrad, M., WZB Data Science Blog, 2017, November. Topic Model Evaluation in Python with tmtoolkit. Available at: https://datascience.blog.wzb.eu/2017/11/09/topic-modeling-evaluation-in-python-with-tmtoolkit/. [Accessed 02 September 2018].

Krestel, R., Fankhauser, P. and Nejdl, W., 2009, October. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems* (pp. 61-68). ACM.

Lettier, 2018, February. Your Easy Guide to Latent Dirichlet Allocation. Available at: https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji

Li, W. and McCallum, A., 2006, June. Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd international conference on Machine learning (pp. 577-584). ACM.

Long, R., Wang, H., Chen, Y., Jin, O. and Yu, Y., 2011, September. Towards effective event detection, tracking and summarization on microblog data. In International Conference on Web-Age Information Management (pp. 652-663). Springer, Berlin, Heidelberg.

Lozano, M.G., Schreiber, J. and Brynielsson, J., 2017. Tracking geographical locations using a geo-aware topic model for analyzing social media data. Decision Support Systems, 99, pp.18-29.

Phuvipadawat, S. and Murata, T., 2010, August. Breaking news detection and tracking in Twitter. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 3, pp. 120-123). IEEE.

Popescu, A.M. and Pennacchiotti, M., 2010, October. Detecting controversial events from twitter. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1873-1876). ACM.

Popescu, A.M., Pennacchiotti, M. and Paranjpe, D., 2011, March. Extracting events and event descriptions from twitter. In Proceedings of the 20th international conference companion on World wide web (pp. 105-106). ACM.

Richardson, G., Bowers, J., Woodill, A., Barr, J., Gawron, J. and Levine, R., 2014. Topic Models: A Tutorial with R. International Journal of Semantic Computing. 08. 85-98. 10.1142/S1793351X14500044.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M. and Both, A., 2014. Evaluating topic coherence measures. arXiv preprint arXiv:1403.6397.

Rouse, M., Search Business Analytics, 2017, June. What is social media analytics? - Definition from WhatIs.com. Available at: https://searchbusinessanalytics.techtarget.com/definition/social-media-analytics. [Accessed 19 August 2018].

Russell, M., 2013. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. 2nd ed. UK: O'Reilly Media, Inc.

Sakaki, T., Okazaki, M. and Matsuo, Y., 2010, April. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (pp. 851-860). ACM.

Saravanakumar, M. and SuganthaLakshmi, T., 2012. Social media marketing. Life Science Journal, 9(4), pp.4444-4451.

Sokolova, M., Huang, K., Matwin, S., Ramisch, J., Sazonova, V., Black, R., Orwa, C., Ochieng, S. and Sambuli, N., 2016. Topic Modelling and Event Identification from Twitter Textual Data. arXiv preprint arXiv:1608.02519.

Stieglitz, S., Dang-Xuan, L., Bruns, A. and Neuberger, C., 2014. Social media analytics. Business & Information Systems Engineering, 6(2), pp.89-96.

Stieglitz, S., Mirbabaie, M., Ross, B. and Neuberger, C., 2018. Social media analytics–Challenges in topic discovery, data collection, and data preparation. International Journal of Information Management, 39, pp.156-168.

Tang, K., Personalized Emotion Classification with Latent Dirichlet Allocation.

Wen, T.H., Kung, P.H. and Su, P.H., Analysis of combining Topic model, Sentiment, Geolocation information approaches on Social Network.

Yang, S.H., Kolcz, A., Schlaikjer, A. and Gupta, P., 2014, August. Large-scale high-precision topic modeling on twitter. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1907-1916). ACM.

Zeng, D., Chen, H., Lusch, R. and Li, S.H., 2010. Social media analytics and intelligence. IEEE Intelligent Systems, 25(6), pp.13-16.

Zhao, Y., Zhang, K., Zhang, H., Yan, X. and Cai, Y., 2017, December. Hot topic detection based on combined content and time similarity. In Progress in Informatics and Computing (PIC), 2017 International Conference on (pp. 399-403). IEEE.

Zhu, B., Yu, Y., Li, C. and Wang, H., 2017, December. Research and implementation of hot topic detection system based on web. In Computer and Communications (ICCC), 2017 3rd IEEE International Conference on (pp. 1504-1509). IEEE.

Zou, L. and Song, W.W., 2016, July. LDA-TM: A two-step approach to Twitter topic data clustering. In Cloud Computing and Big Data Analysis (ICCCBDA), 2016 IEEE International Conference on (pp. 342-347). IEEE.