Application of short text topic modelling techniques to Greta Thunberg discussion on Twitter

MSc Research Project
Masters in Data Analytics

Sean Dingemans
Student ID: x18199089

School of Computing
National College of Ireland

Supervisor: Dr Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sean Dingemans……………………………………………………………….……………
Student ID:	X18199089……………………………………………………………………………………
Programme:	Masters in Data Analytics **Year:** 2020…………….
Module:	MSc Research Project …………………………………………………….….….….……
Supervisor:	Dr Catherine Mulwa …………………………………………………………….….……
Submission Due Date:	17 August 2020………………………………………………………………………….………
Project Title:	Application of short text topic modelling techniques to Greta Thunberg discussion on Twitter………..………………………………………………
Word Count:	10,850……………………… **Page Count**… 26……………………………….……….

I hereby certify that the information contained in this (my submission) is information pertaining to the research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ………………………………………………….………

Date: 17 August 2020………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on the computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Application of short text topic modelling techniques to Greta Thunberg discussion on Twitter

Sean Dingemans
Student ID x18199089

Abstract

Twitter has become an essential medium for probing differing views on issues within society. One such issue is climate change, which is of interest to social scientists. Topic modelling is a way to discover such discussion and viewpoints. Latent Dirichlet Allocation (LDA) is the most widely used topic modelling technique. However, there are limitations for LDA with texts typically less than fifty words, where it suffers from poor characterisation. Consequently, six novel topic modelling algorithms were evaluated against LDA to assess their performance. Topics around climate change activist Greta Thunberg were to be examined as a secondary objective. The collected data comprised of Tweets centring on her United Nations Climate Action speech. Topic allocations for the topic modelling algorithms were evaluated with a novel combination of classification recall and coherence scores. The algorithms Word Network Topic Model (WNTM) and Biterm Topic Model (BTM) were found to have the best overall performance. These novel algorithms will be of great interest to social scientists and marketing companies who would like to probe the discussions on Twitter better. Discussion around Greta Thunberg was found to be polarised.

1 Introduction

Topic modelling is a text summarisation technique used to extract hidden themes from collections of texts known as corpora in the form of topics. The most well-known topic modelling methodology Latent Dirichlet Allocation (LDA) is finding increasing application in the social sciences and has been used to study climate change in collections of text documents [1], [2]. The subject of climate change is becoming an ever-increasing topic of discussion of everyday life amongst individuals with many people strongly acknowledging or strongly denying the concept of climate change [3], [4]. An aspect of climate change opinion that has not been adequately addressed in the literature is the analysis of social media conversational themes concerning the implementation of climate change policies.

An interesting individual to study with regards to the climate change discussion would be Greta Thunberg. Greta is a unique grassroots individual who has come to represent the frustration of climate change inaction for millions of individuals. She brings a new dimension to the discussion of climate change, mainly through the medium of *Twitter*[1] – the ability to mobilise the youth and hold political leaders to account. With such a strong influence on school children, her impact can be long-lasting considering that these children will reach voting age in the not too distant future. Consequently, climate change can be expected to have a more substantial influence on local and regional politics. Policy and

[1] https://about.twitter.com/

action for climate change have thus become increasingly important for governments around the world to address.

A variety of topic modelling methodologies were utilised to extract the themes of discussions on Twitter about Greta in relation to her Climate Action speech at the United Nations[2] in September 2019.

1.1 Motivation and Background

Short Text Topic Modelling (STTM) techniques have not, as yet, been evaluated against each other thoroughly on a Twitter dataset of considerable size (over 100,000 Tweets). Surprisingly, many online Natural Language Processing (NLP) services, such as Amazon Web Services (AWS) *Comprehend*[3], and Google *NLP*[4] have not implemented STTM techniques into their current NLP suites, and are still relying on traditional LDA modelling methods. The modelling technique limitations are a problem, as LDA does not adequately cater for short texts (less than fifty words). Additionally, none of the short text topic modelling algorithms has been applied in a relevant social science domain. These algorithms thus deserve a detailed evaluation on subject orientated data, such as climate change, to evaluate their effectiveness.

Twitter is becoming an increasingly important tool to probe people's opinions on current issues within the world. Twitter also has many advantages public opinion surveys [4], as more people are reached on a broader scale. It allows for the examination of expression, free of polling constraints – often in a conversational manner [5]. Greta Thunberg's most important medium for climate change communication is Twitter. Consequently, much invaluable discussion on climate change has taken place on Twitter around Greta.

1.2 Research Question

Twitter is an invaluable platform to gauge the opinions of society. It would be invaluable to grassroots movement stakeholders such as climate change action groups like *Sunrise Movement*[5] and *350.org*[6], to reliably classify Tweets in specific themes in an objective manner that can be consistently related to the contents of a corpus of aggregated Tweets.

There were two main goals for this project. The first goal was addressed as a primary Research Question with a focus on evaluating the relevant topic modelling methodologies. The second goal was addressed as a Sub-research Question, focussing on the exploration of the themes within the Twitter data, once an ideal topic modelling technique was identified.

> ***RQ:*** *Can short text topic modelling techniques (DMM, LF-DMM, GPU-DMM, BTM, and WNTM) outperform LDA, to reflect the properties of the document corpus on Twitter appropriately?*

---

[2] https://www.un.org/en/climatechange/un-climate-summit-2019.shtml
[3] https://aws.amazon.com/comprehend/
[4] https://research.google/research-areas/natural-language-processing/
[5] https://www.sunrisemovement.org/about
[6] https://350.org/

Sub-RQ: *What are the common themes of discussion around Greta Thunberg concerning her 23 September 2019 address*[7] *to the United Nations Climate Action Summit?*

1.3 Research Objectives

For the research questions to be successfully addressed, the following research objectives were identified, as defined in Table 1 below.

Table 1: Research Objectives

Objective	Description	Evaluation Methodology
O1	A thorough examination of the relevant literature to motivate the current work and identify the research gaps;	Critical thinking;
O2	Identification of relevant timeframe for the collection of the Twitter data and the collection of the data itself;	Find timeframe related to relevant event and examine Twitter traffic around the event;
O3	Pre-processing of the collected data;	Follow guidelines from relevant literature on data preparation, make decisions based on data exploration;
O4	Parameter selection for models;	Preliminary of output from model tuning, evaluation of rules of best practice from literature;
O5	Implementation of topic modelling algorithms (LDA, DMM, LF-DMM, GPU-DMM, BTM, and WNTM);	Classification recall, analysis of sparsity, coherence testing, and distance metrics, token evaluation. chi-squared statistics;
O6	Evaluations of the output of implemented topic models;	
O7	Evaluation of the content of topics surrounding the climate change discussion.	Time series plots, the relation of discovered topics to available literature, chi-squared significance values.

1.4 Contribution to Body of Knowledge

The significant contribution to the scientific literature was the critical evaluation of STTM techniques against traditional LDA on gathered Twitter data coinciding with a well-publicised event of international relevance. The event of note was the United Nations (UN) Climate Action Speech, given by Greta Thunberg on 23 September 2019. The research entailed the identification of the optimum topic modelling methodology and the discovered relevant trending topics.

The structure of the research report was subdivided into five different focus areas in the sections following. Section 2 is an evaluation of related academic work, while section 3

[7] https://text.npr.org/s.php?sId=763452863

discusses the motivation for the adapted research methodology approach and the modifications to the design specifications. Section 4 defines the implementation of six different topic modelling methodologies used in the research project. Section 5 then discusses the evaluation, including the assessment of the evaluation metrics, and the assessment of the topics generated by the topic models around the climate change discussion associated with Greta Thunberg. Section 6 presents the conclusions and provides insight into future work. The evaluation of the academic work on related prior work is discussed in the section following.

2 Related Work

The literature review initially focuses on the previous work, examining various aspects of climate change discussion on Twitter. The review then discusses traditional topic modelling – namely LDA, its aims and how it is used, before discussing relevant applications of LDA in communications studies. There is further discussion about the issues associated with traditional topic modelling for short text documents, and how alternative topic modelling techniques can address these issues. The evaluation metrics of the topic modelling methodologies used were discussed before concluding with a discussion on identified gaps in the previous research.

There are six sub-sections in the literature review, namely Previous Climate Change Related Twitter Studies, Traditional Topic Modelling, Applications of Topic Modelling in Communication Studies, Short Text Topic Modelling, and A Critical Revie of Evaluation Techniques and Metrics, and a Conclusion with Identified Gaps, which are presented in the sections following.

2.1 Previous Climate Change Related Twitter Studies

Previous studies have varied in their focus, examining sentiment analysis [5]–[8], the structure of user interaction [9], [10], time series analysis of Twitter traffic [5], [11], [12], the analysis of Hashtags and words used in Tweets, and their associated metadata, discussion flow and content [4], [7], [13], or a combination of the various approaches above [5], [6], [8].

The terminology for the global greenhouse effect on Twitter has changed over time, with the studies of [6], [8] illustrating a gradual move away from the term *'global warming'* to the term *'climate change'*. The term *'global warming'* seems to be now associated with the negative sentiment, more so than *'climate change'* – despite both terms describing the same concept. The relevancy of a search term must thus be carefully considered when collecting the data. The authors of [8] also saw a change in the polarity of the language used over time, with the associated discussion becoming more divisive and partisan.

Brief public events with high media exposure tend to attract considerable Twitter traffic [5], [11], [14]. Examination of the volume of traffic related to the event is usually tracked via Hashtags and pertinent keywords. Often, it was found that these events have a short-term effect, with the Twitter volume related to an event usually tailing off after a few days. For events that lasted for a few days, [14] observed a plateau of activity for high profile social events – such as conferences, gala events and ceremonies. The Twitter content was found to relate to what was discussed at the event, as noted by [7], [11]. The findings of these studies were based on the relevant term frequencies without consideration of in document word co-occurrences – thus potentially missing out on nuanced themes within the textual data. These research findings suggest that people only tend to post topics in response to defined topical short-term events – possibly to share their thoughts at a time when others are showing an interest in a particular event. The study by [11] validates this claim, by examining the relative search volume (RSV) of Google searches of the 2016 Academy Awards[8] versus the Twitter RSV volume of the 2016 Academy Awards. A strong correlation exists between the RSV's on both platforms for the event.

Additionally [12] found that the weekly Tweet volume associated with climate change correlated to the volume of articles published by major news media outlets – over a third of all Tweets in the dataset had a hyperlink to published external media content. Thus, it seems that traditional established media has a significant influence on the Twitter volume (mainly through Retweets) and possibly even the Tweet content. Several influential climate action related accounts have a strong influence on Twitter volume. Fifty per cent of all Retweets are related to just fifty accounts – again indicating the centralisation of the conversation by a few active participants.

With further regards to network centralisation – when there is interaction on Twitter concerning climate change, interaction is commonly grouped into various communities on the platform. Generally, individuals within a 'community' tend to share more like-minded views and to share the positive in-group sentiment. These networks were constructed on a Retweet or follower-follower basis. For both of the produced network graphs in studies [9] and [10], the nature of outgroup interactions depended on the ideological stance of the other community, tending to be more negative if the ideological stances differ. Additionally, [9], [10] found that sceptic users were more proactive in expressing negative sentiments to outgroup activist users. Activists frequently clustered together, with more relevant discussion found in their respective communities.

A study by [13] that used topic modelling implemented Latent Semantic Analysis (LSA). The analysis of the content from this study was more insightful in comparison to other studies, with three thematic areas being found within the texts – call for action, climate policy, and consequences. Discussion of these findings, though, was limited, and the data collected was not directly related to any specific short-term event.

While these studies have proven insightful, they have not thoroughly examined potential themes and topics running through the content. Topic modelling helps provide a clearer insight into a Tweet's content. What exactly is discussed in a Tweet – can the Tweet content be appropriately generalised to an underlying theme or topic within the aggregated data? For this aspect to be examined, various topic modelling techniques need to be utilised.

2.2 Traditional Topic Modelling

Currently, the most widely used topic modelling technique, serving as a benchmark for other techniques, is LDA. LDA is an unsupervised Bayesian clustering algorithm based upon a Dirichlet Allocation. In LDA, a topic, or multiple topics, are assigned to each document.

[8] https://www.oscars.org/oscars/ceremonies/2016

The process begins with a uniform multinomial distribution, or Dirichlet prior, where the number of specified *'k'* outcomes represent themes within the corpus [15]. LDA uses two Dirichlet priors – one to generate a topic-document matrix and the other to generate a topic-document matrix. The parameters *alpha* and *beta* can be used to control the sparsity and skewness of the resulting two discrete distributions. Sparsity and topic imbalance often occurs naturally within corpus collections and are thus essential to model [16]. Initially, the algorithm assigns a discrete stochastic distribution of topics to each document in the corpus. These discrete distributions are, in turn, drawn from the Dirichlet distributions, helping to parameterise the drawn distributions. Through an iterative process of examining words in the documents, the model revises the initial distribution of the two discrete distributions *theta* and *phi* by solving the current joint probability distribution of the parameters within the algorithm.

The discrete distributions *theta* and *phi* are multinomial and create a Joint Probability Distribution that can only be solved through approximation. The most common and popular approximation methodologies for this process of inference is Gibbs Sampling (GS), and Expectation-Maximization (EM) based inference [17]. Gibbs Sampling essentially assumes some independence of the variables in the joint probability distribution, and iteratively uses individual conditional distributions to approximate the true joint probability distribution. The joint probability distributions of these matrices are approximated so that they are guaranteed to converge, unlike EM. Following on from the LDA overview, topic modelling application in social science studies can now be examined.

2.3 Applications of Topic Modelling in Communication Studies

In the examination of political discussion by general society, [18] conducted both traditional opinion polling and Twitter discussion analysis to explore issues of importance to the German public for the 2013 Bundestag Elections. LDA was used for Twitter topic modelling. In addition to finding the topics defined by participants from the survey, LDA found many additional topics on Twitter. The authors of [18] found that Twitter was a more relevant platform for tracking discussion around societal events, in comparison to Facebook.

Previous applications of LDA topic modelling to examine environmental discussion have been found in [1], [2], [19]. The authors of [1], [2] selectively studied articles and posts released by United States conservative think tanks. Both studies aimed to find hidden themes within climate change centric articles. The findings in the studies suggested a gradual shift in the tone of themes from being more policy-driven to more science-driven – these being the result of focussed efforts made to discredit any scientific findings providing evidence of the human impact on climate. Personal attacks became more frequent when prominent individuals, like Al Gore, began to speak out on climate change. The same can be expected to be seen for Greta Thunberg. The change in more policy-related articles (articles aimed more at the political elite and their staff) to more science sceptic articles (articles aimed at a more general audience) highlights a general feeling of concern amongst fossil fuel dominated conservative industries. In response to ordinary individuals and scientists voicing concern about the human impact on climate, the focus of these think tanks changed from 2007. According to [2], [19], the general goal was to confuse the public, in order to weaken climate change grassroots movements. The authors of [2] found an increased amount of funding for conservative think tanks that published climate change sceptic news articles, which could give weight to their claims. In studying blogs discussing climate change, [19] found further evidence of ideological polarisation around the issue of climate change. It is of interest to see if these polarisation trends can be seen thematically on Twitter. In order to extract potential themes, the selection of an appropriate topic model algorithm is required.

2.4 Short Text Topic Modelling

A common problem though with LDA topic modelling is the tendency to struggle with short text documents like Tweets. This problem is due to a very high level of sparsity in short texts, where the vocabulary of the corpus is orders of magnitude larger than the average number of words per document [20]. The latent topic distribution is also naturally skewed on social media platforms. Additionally, the use of LDA on social media typically results in poorly defined topics, which tend to be more generalised[21].

Few communications studies have utilised alternatives to LDA for their domain analysis. The studies of [13], [18] each only partially examined one alternative to LDA. There are, however, a variety of technical papers that have found alternatives to traditional LDA, or altered the appropriate algorithms to handle short sparse texts. One such modification is to aggregate Tweets by username into a single pseudo-document [22], [23]. This assumption, however, results in the gross loss of resolution of the individual Tweets [24]–[26].

Consequently, researchers have looked to other related algorithms for shorter texts. One such related algorithm is Dirichlet Multinomial Mixture (DMM). The algorithm for DMM is similar to that of LDA, but assumes that each document only has one topic.

DMM was modified by [25] to use collapsed Gibbs Sampling. Applications of this modified DMM are superior to that of traditional LDA [25]–[27]. In work undertaken by [27], [28], DMM was further modified by incorporating word embedding features into the model. Using the English *wikicorpus*, word vectors were trained to capture word context and word relatedness. By doing so, related terms of similar semantic meaning are more likely to be grouped. The Latent Features (LF), or word embeddings, are then linked to the DMM model with either Generalised Poly Urn (GPU) sampling [27] or using Bernoulli sampling [28]. Both Generalised Poly Urn Dirichlet Multinomial Mixtures (GPU-DMM) and Latent Feature Dirichlet Multinomial Mixtures (LF-DMM) use the trained word embedding as a cross-reference for weight allocation of the topic-word matrix. The embeddings give words higher weights if they are contextually and syntactically related to other words in a topic. In comparison to DMM, GPU-DMM and LF-DMM are computationally expensive, as they need to make more approximations with Gibbs Sampling inference, due to the coupling of the word vectors with DMM [27].

The authors of [29] proposed a Biterm Topic Modelling (BTM) method, an extension of non-negative matrix factorisation. BTM makes use of a biterm – unordered pair of words occurring in the same document. Instead of generating a topic-document distribution, BTM generates a biterm-topic distribution. The biterm-topic distribution allows for a combination of words to be modelled to topics, instead of treating each word separately. By not accounting for document generation, results from BTM need to be inferred back onto the corpus documents. For this inference, DMM with Gibbs sampling is used.

The final model discussed is the Word Network Topic Model (WNTM), developed by [21]. Like BTM, WNTM also forgoes document generation. Instead, a word co-occurrence network was generated where there are weighted vertices between words that share the same context. Context, in this respect, can refer to a document, or a sliding window approximation for the document. The newly generated weighted graph can then be used for Gibbs Sampling to generate topic distributions for the words and the pseudo-documents. Finally, LDA with Gibbs sampling is needed to infer the discovered topics back onto the corpus documents.

2.5 A Critical Review of Evaluation Techniques and Metrics

Commonly, distances are measured amongst produced topics – as related topics would tend to be closer to one another. One popular measure for distances is the Hellinger Metric[9]. Once Hellinger Distances are generated, the results can be visualised and evaluated graphically using Multi-dimensional Scaling (MDS) plots [1], [13].

Another popular metric for topic validation is topic coherence. Often, topic coherence scores are used to see if a topic makes sense. Coherence scores can be used to analyse the quality of the generated topics, using a metric C_v, due to C_v displaying the strongest correlation to human coherency scores [30]. The C_v methodology examines word pair correlation, the probabilities of the word pairs, significance testing of the word pair correlations, and the aggregation of the individual scores into an overall coherence score.

Accuracy and classification recall of topic assignments can also be used to validate how well the generated topics relate to the documents themselves. With regards to using features for classification in text analytics, topic distributions per document were used by [21], [27]–[29] to predict the assigned topic label. In contrast, studies such as [31], [32] instead used the corpus word vectors as features for the classification of labelled documents in the corpus. Surprisingly, feature word vectors have not been previously used for classification evaluation in topic modelling. The words are inherent features of the documents themselves (in comparison to topic proportions, which are only generated after the topic modelling).

Consequently, their applicability should be investigated further. Typically, the most suited algorithm for text analytics topic modelling is the Support Vector Classifier (SVC) [31], [32]. SVC is also used in the topic modelling studies by [21], [27]–[29], albeit with topic distribution per document and pre-defined labels.

2.6 Conclusion of Literature Review and Identified Gaps

So far, no comprehensive study on Twitter data has been undertaken to evaluate the effectiveness of novel short text topic modelling algorithms in comparison to LDA. STTM algorithms also need to be evaluated against each other. For evaluation, no papers have examined sparsity of the algorithms, made use of word feature vectors for classification or thoroughly utilised the coherence metric C_v. With regards to investigating climate change discussion on Twitter, no study has made any adequate use of topic modelling. In the next section, the research methodology of the project is described.

[9] https://www.jstor.org/stable/20142523?seq=1

3 Research Methodology

The research methodology and the approach taken in this study are discussed under two headings, namely: Topic Modelling Methodology, and Project Architecture in the sections following.

3.1 Topic Modelling Methodology

The Cross-industry Process for Data Mining (CRISP-DM[10]) was chosen and adapted as the topic modelling methodology template for the project. In the project, the topic modelling methodology considered softer, less rigid technical requirements – namely business understanding and value. The presentation of the results and findings was considered as the deployment phase. The adapted methodology used was cyclical and thus allowed for each step of the cycle to be revised logically. The data mining methodology process flow for the topic modelling project is presented in Figure 1 below.

The project understanding encompasses the elucidation of the scope, aims and rationale for the project. Data was then collected and explored, based on the examination of the number of Tweets over a defined period and the content of the selected Tweets. Following the exploration of the data, adjustments could be made to the selected period and search criteria, if necessary, and then the Twitter scraping collection process could be redone. Text processing could then be undertaken, which involved the removal of stop words, Retweets, and lemmatisation. Subsequently, further data exploration of the Tweets was undertaken to make sure the data was adequate for model training, and that the modelling parameters were relevant for the model runs. After the topic model was trained, the results were examined in the model evaluation phase, by assessing the topic model output. If the results were inadequate due to a lack of data, then more data could be collected, or the scope readjusted. Once key insights were objectively gained, then findings were presented in an appropriate format in the results presentation phase.

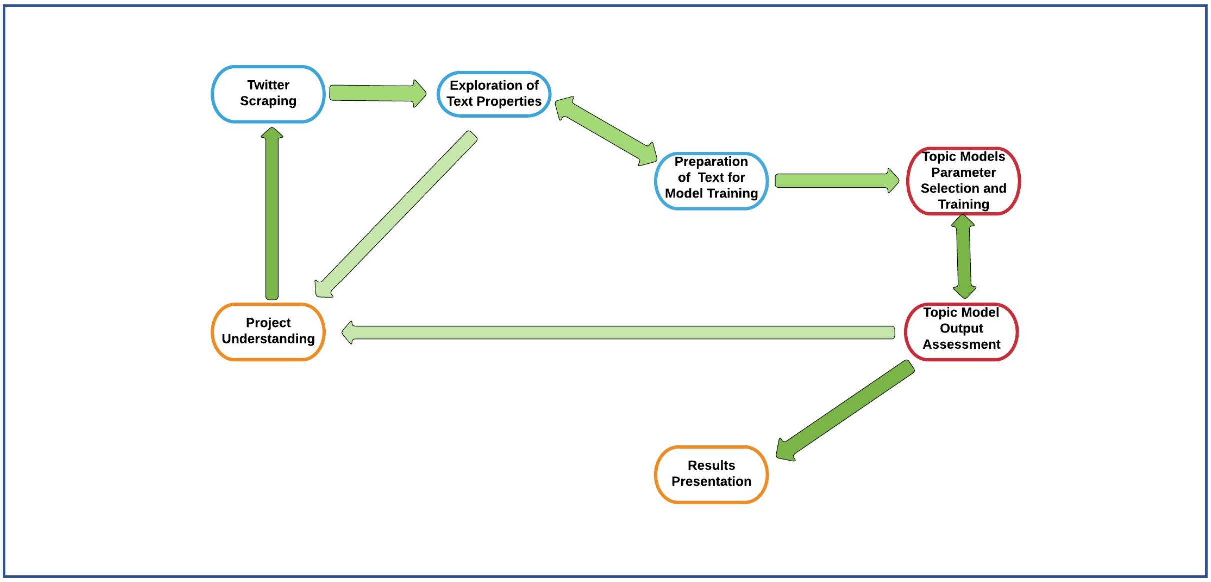

Figure 1: Topic Modelling Project Methodology.

[10] https://www.sv-europe.com/crisp-dm-methodology/

The key aspects of each element of the topic modelling methodology for the project are summarised below:

Project Understanding: The elucidation of project requirements and feasibility of the implementation;

The Data Collection: The extraction of keyword related data (Tweets containing '*greta*') from Twitter over a defined period (20 September 2019 until 30 September 2019);

Data Exploration: Examination of corpus vocabulary, document lengths, Twitter volume over various time-periods;

Data Pre-processing: Stop word, emoji, punctuation removal, lower case conversion, removal of frequent and infrequent words (over 100,000 and under six words), removal of numbers and short words, removal of stop words, removal of Retweets, creation of external word vectors, and removal of short documents (less than three words);

Model Training: Model parameter calibration, feeding of the data into algorithms, troubleshooting the implementation of the algorithms, (like memory allocation problems);

Model Evaluation of the Results: The examination of relevant topic model metrics (such as document-topic distribution, coherence scores and classification recall scores) and the evaluation of the topics produced by the various topic modelling methodologies;

Deployment of Results: Create a visual presentation of the results and write up the conclusions.

The way the topic modelling methodology integrates into the project architecture is discussed in the next section.

3.2 Project Architecture

The stages in the project architecture were broadly grouped into three tiers: a Collection and Processing phase, a Modelling, Training and Evaluation phase, and a Results and Conclusion Presentation phase, as is presented in Figure 2 below.

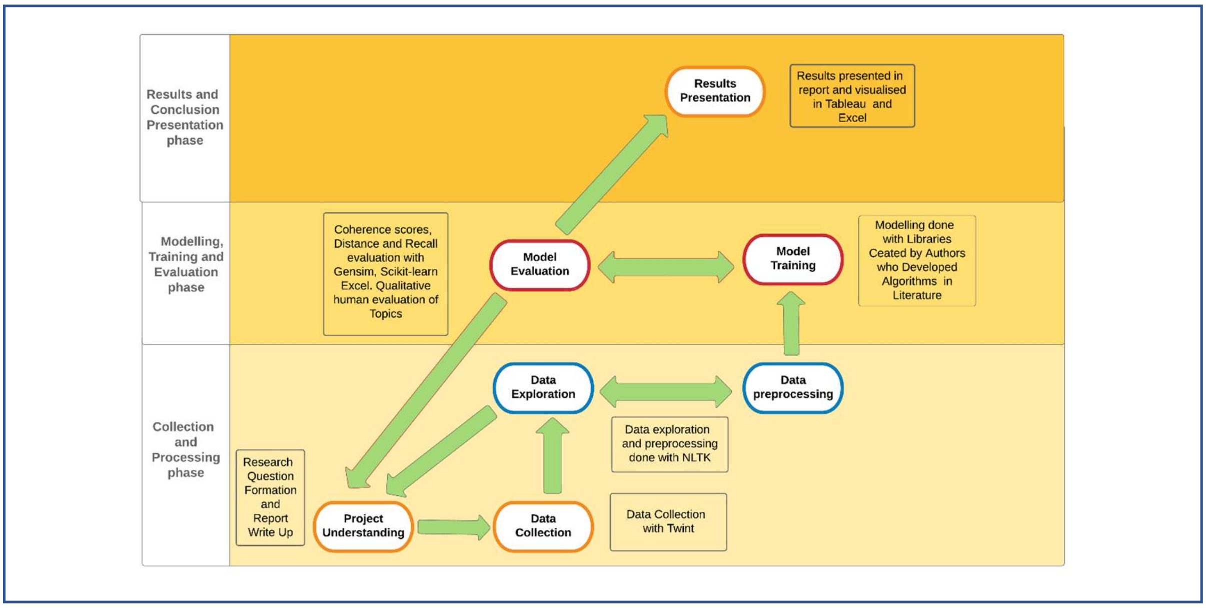

Figure 2: Three-tiered approach for Project Architecture.

The project architecture allowed for flexibility within the different process stages of the research project. The research methodology was implemented in a cyclical approach to refine the most appropriate execution and evaluation of the topic modelling techniques. The project architecture allowed for the sequential and systematic implementation of the project. In the following section data collection, pre-processing, model implementation and results from the evaluation are in the section following.

4 Implementation

The Implementation Phase of the project is discussed under the Data Collection and Pre-processing requirements and Topic Model Parameter Selection and Configuration in the sections following.

4.1 Data Collection and Pre-processing

A popular Open Source Intelligence (OSINT) tool called Twint[11] was used to collect the Tweets. A simple search term '*greta*' was used to scrape Tweets with the term '*greta*'. Some Tweets in threads with '*greta'* in a different context were also inadvertently captured. It was also specified in the Twint command to collect only English language Tweets.

Data from Twitter was extracted over ten days from 20 September to 30 September 2019, to examine the Twitter coverage of Thunberg's UN Climate Action Speech on 23 September 2019. This particular speech provides an elevated platform. The impassionate and scolding nature of the speech itself led to immense press coverage in comparison to her other speeches on climate change activism [14]. The event also saw a surge in the number of her Twitter followers [14].

Duplicated Tweets were removed. Regular expression patterns were used to remove all URLs, Hashtags and user handles. Visual examination was done on the collected Tweets to check that the obtained Tweets were in English. All non-Latin characters and any numerical values were removed from the text, as well as any single letter characters. Text normalisation was then undertaken, with the removal of all punctuation and the conversion of all text to lower case. The text was then tokenised and had a default list of Natural Language ToolKit (NTLK[12]) English stop words removed. After a detailed document length examination, all Tweets with less than three words were removed. In the subsequent examination of the token frequencies, many additional words were added to the stop word list, including words with frequencies over 80,000, words with frequencies under six, nonsensical words, and selectively half of all words with two characters. Most decisions made in the text pre-processing were based on relevant pre-processing exercises from other studies [18], [21], [27] and from the examination of the collected data (see the configuration manual). Upon examination of the corpus, it was decided not to utilise stemming or lemmatisation. The use of lemmatisation would remove the subtleties and context of many words, often normalising words that are best not equated – such as: *children – child;* and *acting – action – act – actress – activist*.

[11] https://github.com/twintproject/twint
[12] https://www.nltk.org/

The corpus comprised a vocabulary of 29,000 words derived from 570,000 documents. The cleaned Tweets were then written to a text file, with one Tweet per line. Minor format adjustments were later applied to the corpus to meet each of the modelling algorithm's input requirements (see the configuration manual).

4.2 Topic Model Parameter Selection and Configuration

The techniques, architecture and framework that underlie the implementation of the topic modelling techniques and the associated requirements are identified and presented in the section below.

Six algorithms discussed in the literature were utilised and compared to one another. Five libraries were used for the six algorithms LDA, DMM, LF-DMM, GPU-DMM, BTM and WNTM (see reference list in the configuration manual).

The core parameters for all the topic modelling methods are *alpha, beta* and *k*. *Alpha* controls the sparsity of topic distribution, *beta* determines the weightings of words per topic, and *k* is the number of designated topics to be generated from the corpus.

Experimentation of *alpha:* The parameter *alpha* (α) was set to an intermediate value of 0.10 after testing and reviewing generated topics from the recommended *alpha* equal to 50/k [16] and *alpha* at 0.01. The distribution of topics from *alpha* equal to 50/k was deemed too uniform and thus artificial. Corpora comprising short texts often contain an imbalanced distribution (or sparsity) of topics [21], [33]. Lowering *alpha* to 0.01 generated too much sparsity, with 30 topics in the collection not having any associated documents. Setting *alpha* to 0.10 not only allowed all the topics to have allocated documents but also created more interpretable topics. Reviewing the literature [33], [34] showed that lowering *alpha* below 50/k was often necessary for corpora of short texts, due to their tendency to be sparse (see the configuration manual).

Experimentation of *beta:* The parameter *beta* was run at 0.10, 0.05 and 0.01 for the LDA and DMM base models. Upon examination, topics produced with the *beta* at 0.10 and 0.05 generated more topics that were unique, but also less interpretable, thus *beta* of 0.01 was used. A small *beta* of 0.01 was also found to be recommended by the studies [28], [29].

Experimentation of *k*: The parameter *k* was also experimented with at settings of 40, 60 and 80 topics, with the parameter defaults for LDA and DMM based on the studies by [21], [25], [27]–[29]. When *k* at 60 topics was compared to *k* at 40 produced topics, more informative results, with some unique, rare topics were discovered. Using *k* at 80 produced topics that were too similar to one another and some were in some respects less sensical. It was thus decided to set *k* to 60 topics, as the corpus was relatively large when compared to the prior applications of the various selected algorithms in the literature [25], [27]–[29].

Experimentation with the number of iterations: The output produced for the document topic (*theta*), topic word distributions (*phi*), and associated topic vocabulary showed negligible differences after 1,500 iterations, and the generated output was thus deemed to have converged. Separate runs with 2000 iterations showed similar and consistent results. Consequently, the number of iterations for each modelling method was set at 2,000.

In addition to the parameters above, word embeddings were prepared for LF-DMM and GPU-DMM. The word embeddings were trained on the dataset itself, due to many common and unique terms of the corpus not being found in existing pre-trained word embeddings [35].

The generated word embeddings were then incorporated and parameterised into the topic models, based on the specifications in the studies by [27], [28].

Experimentation with skip-gram model window size: A skip-gram model was used with a window size of 10, with an average document having a length of 10.5 words. A skip-gram model was chosen, as opposed to a continuous bag of words (CBOW[13]) model, due to its emphasis on rare words.

Experimentation with WNTM window size: For WNTM, the output needs to be inferred on the actual documents, as the method uses a sliding-window based approach to construct a word network, instead of using the true documents. A window size of 10 was used to build the word network, based on the average document length of 10.5 words. Topics were generated from the pseudo-document word network using either LDA or DMM. LDA is generally used, as the pseudo-documents produced from the word network are quite long and better suited for longer texts. The generated LDA *theta* matrix was subsequently used to infer a new *theta* matrix upon the corpus documents.

4.3 Processing of Topic Model Output Data

Upon completion of the running of each algorithm, the outputted data was processed, primarily on the resulting *phi* and *theta* matrices. Due to the size of the output, processing was undertaken with AWK. The generated word-topic matrix *phi* was transposed for efficiency purposes, before finding the top-twenty weighted words for each topic with an indexed list. The topic allocation per document was defined by the topic label used the topic with the highest average weight per document.

The results of this processing are discussed in the evaluation section following. The evaluation section is split into two sub-sections: one covering the evaluation metrics, including techniques assessing the sparsity, coherence, dissimilarity, classification, chi-squared statistics of the associated topic model output, and the other evaluating the discovered topics relating to the discussion about climate change around Greta Thunberg.

5 Evaluation

The evaluation section focuses on the results of the output from the six different topic models and how best to discriminate between the different topic modelling approaches, particularly in terms of the short text topic modelling of Tweets. Several different metrics were used to evaluate different aspects of the output, namely the sparsity, coherence, dissimilarity, classification and chi-squared statistics. The content of the topics associated with the discussion around Greta Thunberg was used to confirm the validity of the revealed topics and the time and volume related validity of the modelled outputs. The various modelling metrics were combined in an effort to provide aspects of quality control to the validated output. Trends were identified in the different topic modelling approaches, which are discussed in the sections below on Evaluation Metrics.

[13] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0220976

5.1 Evaluation Metrics

The evaluation was done for sparsity, coherency of topics produced, distance metrics, classification, and how well the topics related to their labels. The evaluation was undertaken directly on the quality of the generated topics through distance metrics and coherence scores. In addition, an assessment was carried out on how well the created topics related to the documents from which they were generated, through coherence scores, classification recall and chi-squared tests.

5.1.1 Sparsity

For the topic allocation of documents, the modelling methods vary significantly in their distribution of documents to topics. The LDA modelling methods showed very little sparsity, and this can be seen visually in Figure 3 below. The evenly spread topic document distributions for LDA was viewed with some scepticism. Short text corpora tend to have high levels of sparsity [21], [29], [33]. It was concluded that it was statistically improbable that every topic would be allocated a similar quantity of documents given the diverse nature of the conversation on Twitter.

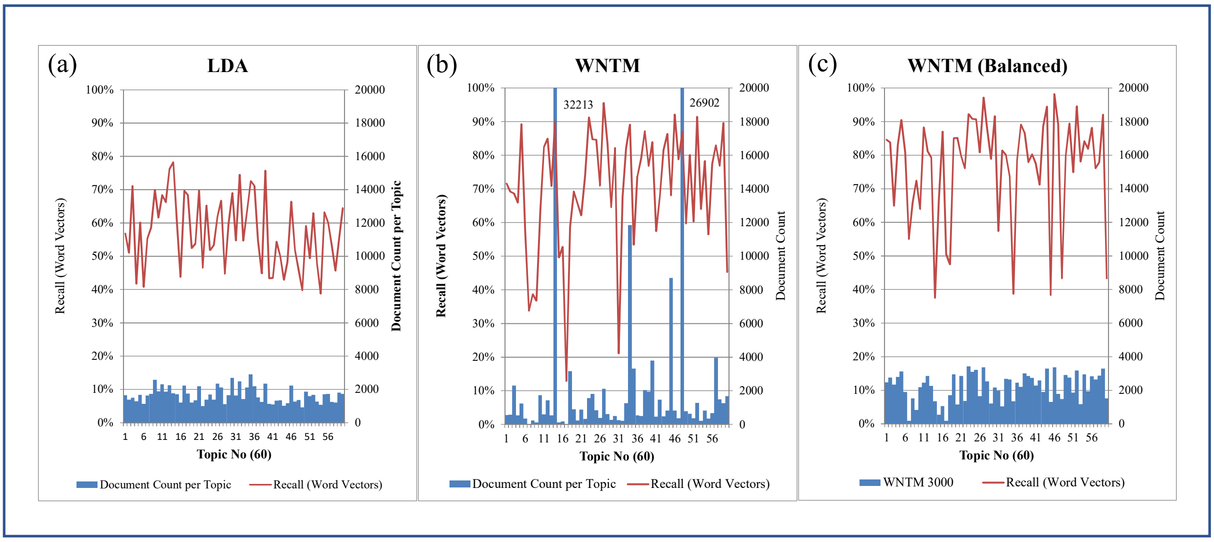

Figure 3: (a) LDA - weak sparsity; (b) WNTM enhanced sparsity, and (c) WNTM Balanced (truncated at 3000) – improved sparsity.

For the DMM modelling methods, good sparsity was found, except for LF-DMM. The reduced level of sparsity in LF-DMM could be due to weak coupling between word embedding and the Dirichlet components of the methodology. The non-generative modelling methods, BTM and WNTM, showed the highest levels of sparsity, with probable over-allocation to a few topics. For example, there was a topic on the Covington incident in both the WNTM and BTM outputs. The term '*covington*' only appears 3,300 times throughout the entire corpus. The WNTM output has 2,100 documents allocated to the Covington topic while in the BTM output 67,000 documents were allocated to the topic – clearly showing over-allocation. Similarly, for WNTM, there were two topics with over 100,000 document allocated with poor coherence values. The over-allocation to such topics could be due to dominant words that are not unique to the generated topic. Examples of such words include the homonyms *'like', 'left',* and *'right'*.

Not all the topics with large document counts can be assumed to be a conglomeration of themes with little inter-relationship. The topic congratulating Greta on her achievements after her UN speech was a coherent topic, where the Tweets are consistent and focus on the generated topic. This particular topic even appears consistently in the time series plots for all six model outputs, although the modelled volumes do not necessarily correlate. Identifying valid topics was thus often tricky, as not all dominant topics with high document counts can be considered valid or invalid. A combination of classification recall results and coherence scores were examined to address the validity and interpretability of the discovered topics.

5.1.2 Coherence

Coherence scores were used for the coherence evaluation to analyse the quality of the generated topics, using the metric $\mathbf{C_v}$ [30]. The library Gensim[14] was used to generate the coherence scores, while the library Scipy[15] was used for the significance testing of the coherence scores. The coherence metric $\mathbf{C_v}$ examines word pair correlation, probabilities of the word pairs and significance testing of the word pair correlation, outputting a value in the range of zero to one.

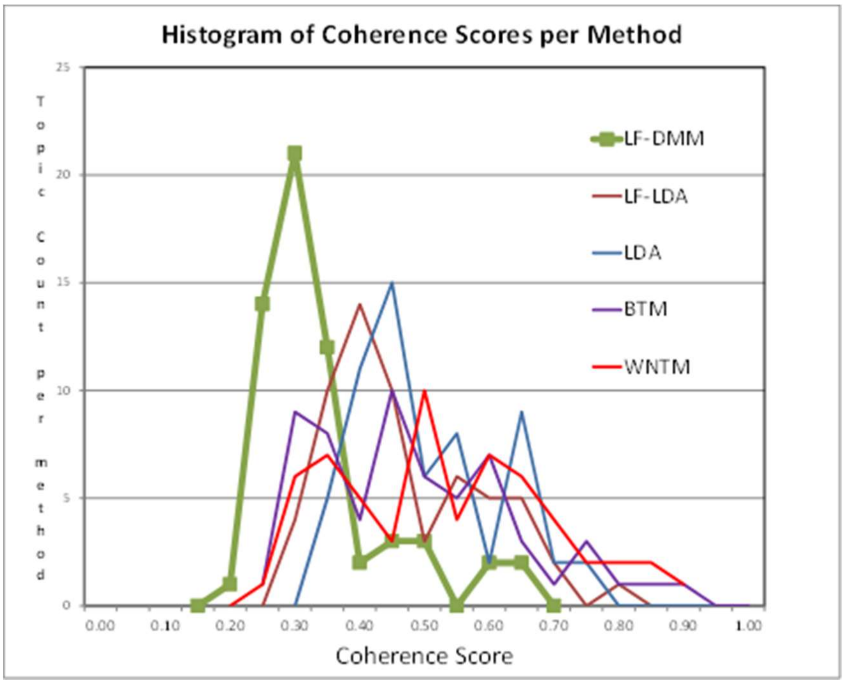

Figure 4: Coherence score graphs for each modelling method.

The distribution of coherence scores per topic modelling method is presented in Figure 4 above. The non-parametric Kruskal-Wallis[16] test was used to examine whether there were any significant differences amongst the distributions of the generated coherence scores per

[14] https://radimrehurek.com/gensim/
[15] https://www.scipy.org/
[16] https://www.researchgate.net/publication/289442433_Methodology_and_Application_of_the_Kruskal-Wallis_Test

topic modelling method. With the null hypothesis of Kruskal-Wallis rejected, adjusted pairwise post hoc testing was undertaken using the Conover-Iman[17] test, due to the test prioritising statistical power. The modelling algorithms BTM, LDA and WNTM were shown to be significantly different were LF-DMM, GPU-DMM and DMM, which have consistently lower. The coupling of word embeddings and associated DMM modelling methods could be a possible reason for the lower scores. From the plot in Figure 4, it seems that the LDA modelling methods are slightly less noisy, having a narrower range in comparison to the non-generative topic modelling methods BTM and WNTM.

5.1.3 Dissimilarity

For distance examination, the Hellinger distance metric was used to find the distance between two documents. The Hellinger scores are symmetric, and take into account the probability weights per word [33]. The library used to generate the distances was Gensim, while the library Scikit-learn[18] was used for the MDS plots. For each topic, the words, together with their respective weightings, were used to generate the Hellinger distance metric. The plots in Figure 5 following show a MDS plot of the distances amongst the topics using their top-20 weighted words. In general, the plots are similar, with a cluster of points in the central part (approximately within ±0.20-dimensional units of the central position) and with more scattered points on the outer parts of the plot.

LDA show a lack of any clustering and has a more widely scattered pattern, as is shown in Figure 5. The LDA methods show the most scattered plots, with uniform distances between each topic, probably due to poor sparsity modelling, while the other topic modelling methodologies tend less to have uniform spacing with more topics clustered closer to the centroid, which was a probable reflection of improved sparsity, as indicated by [33]

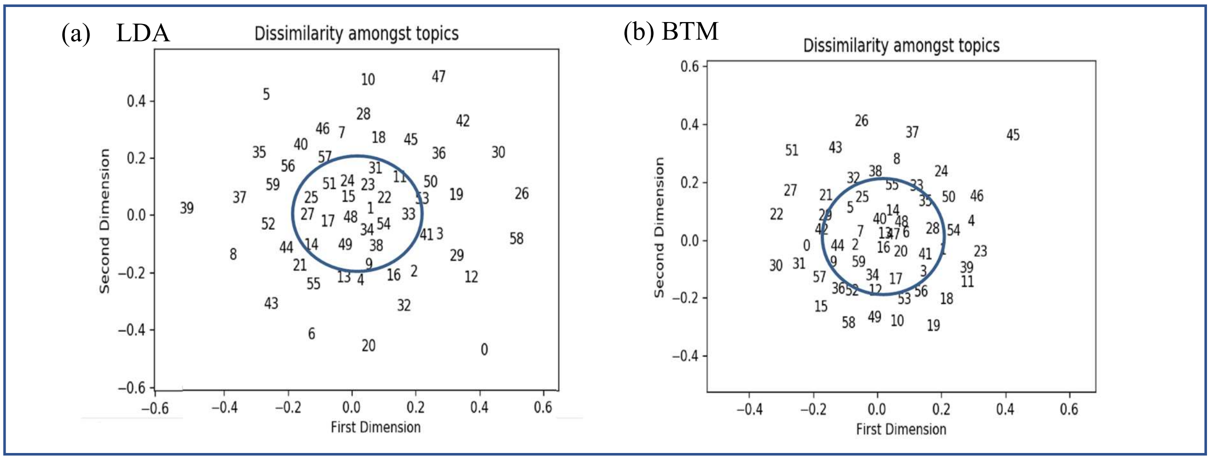

Figure 5: Topic Dissimilarity Plots showing the more diverse scattering of LDA *(a)* compared to the more clustered plot of BTM *(b).*

In a study on long documents by [1], discrete subgroupings of similar topics with similar dissimilarity scores were found in the relatively narrow range of -0.2 to 0.2 units. The dissimilarity scores of topics in this study are higher than those found in the study by [1].

[17] https://cran.r-project.org/web/packages/conover.test/conover.test.pdf
[18] scikit-learn.org

The scattered plots confirm the highly diverse discussion and themes generated in response to a significant and complex topic such as climate change, which has many elements and diverse points of view.

5.1.4 Classification

The classification was undertaken with linear kernel SVC to analyse how well topics related to the documents themselves. SVC was chosen for the classification evaluation over decision tree methods, as it has less of a propensity to overfit, and was frequently used in text classification [31], [32]. To implement SVC, the library Scikit-learn was used.

Data was split 70: 30 for training and testing datasets respectively. A random seed was used for each topic run. Due to the sufficient size of the used dataset, successive runs on each topic model output yielded the same classification accuracy average.

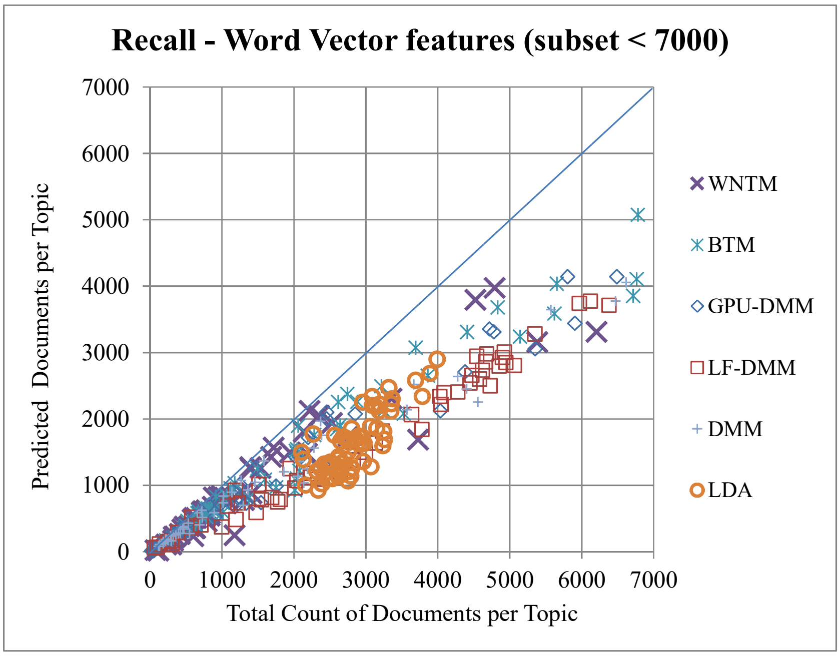

Figure 6: Classification Recall based on Word Vectors
(extracted sub-set plot of documents per topic < 7000).

Classification recall per topic was then examined. Word vectors were used as classification features in separate runs for each modelling method and are displayed in Figure 6. While the topic distribution per document may provide some integrity to the topic allocation per document, pre-defined labels like Hashtags or manual labelling are then required. It was thus deemed more insightful to examine how word features related to the assigned topics, as these are inherent features of the documents. The corpus was converted to Term Frequency–Inverse Document Frequency (TF-IDF) vectors before being utilised in SVC. Results from word features were more discriminating than topic proportions as features. While all the methods show a reduced accuracy for word vector features, LDA showed the second-lowest accuracy at 0.58 despite having balanced topic classes. The word feature vectors thus help show the inherent faults of the LDA methodologies for short text topic modelling.

Table 2: Summary of Classification Recall scores per topic modelling method using word vectors.

Topic Modelling Method	Classification Recall- Word Vectors
WNTM	0.77
BTM	0.74
GPU-DMM	0.67
DMM	0.66
LDA	0.58
LDF-DMM	0.57

The lack of sparsity for the LDA modelling methods can be further seen in Figure 6, with a limited spread for document counts. BTM shows a consistent classification recall across all topic sizes. In contrast, DMM shows a reduction in the classification recall values with larger topic sizes. Deterioration in the classification recall can also be seen for WNTM, GPU-DMM and LF-DMM for topics with less than 2,000 documents. Individual classification recall scores per method are in Table 2 above.

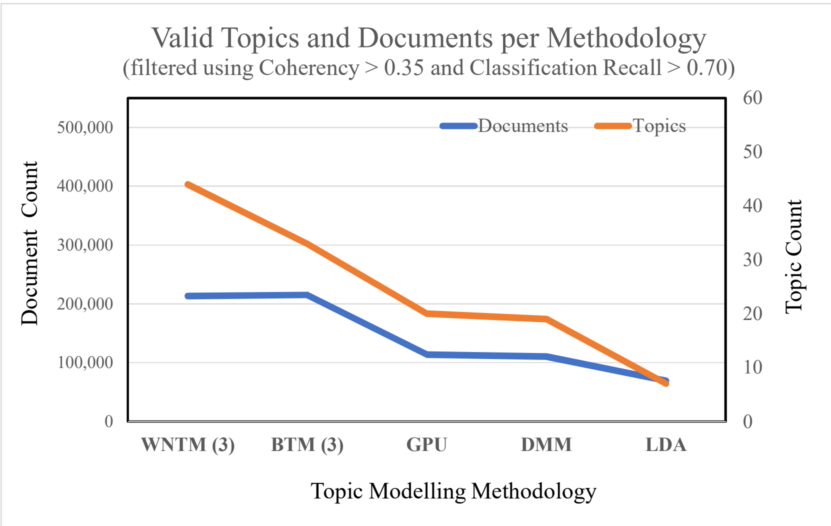

Figure 7: Document and Topic count per Topic Modelling Methodology filtered for coherence scores above 0.35 and classification recall higher than 0.70.

For a topic to be considered valid, the topic should have both a high classification recall score and moderate to high coherency score. Consequently, the word vector topic classification recall scores were examined in tandem with the coherence scores.

Figure 7 shows that when filtering for coherence scores above 0.35 (topics above 0.35 were deemed interpretable) and for classification recall above 0.70 (with the inherently weak over-allocated topics filtered out), BTM and WNTM showed better overall performance.

Upon concluding that BTM and WNTM were the best performing methodologies, further exploration was undertaken to find out how class imbalance affected the results for these two methodologies. Both text modelling methodologies suffered from high class imbalance (sparsity), with smaller classes tending to suffer from poorer classification recall scores.

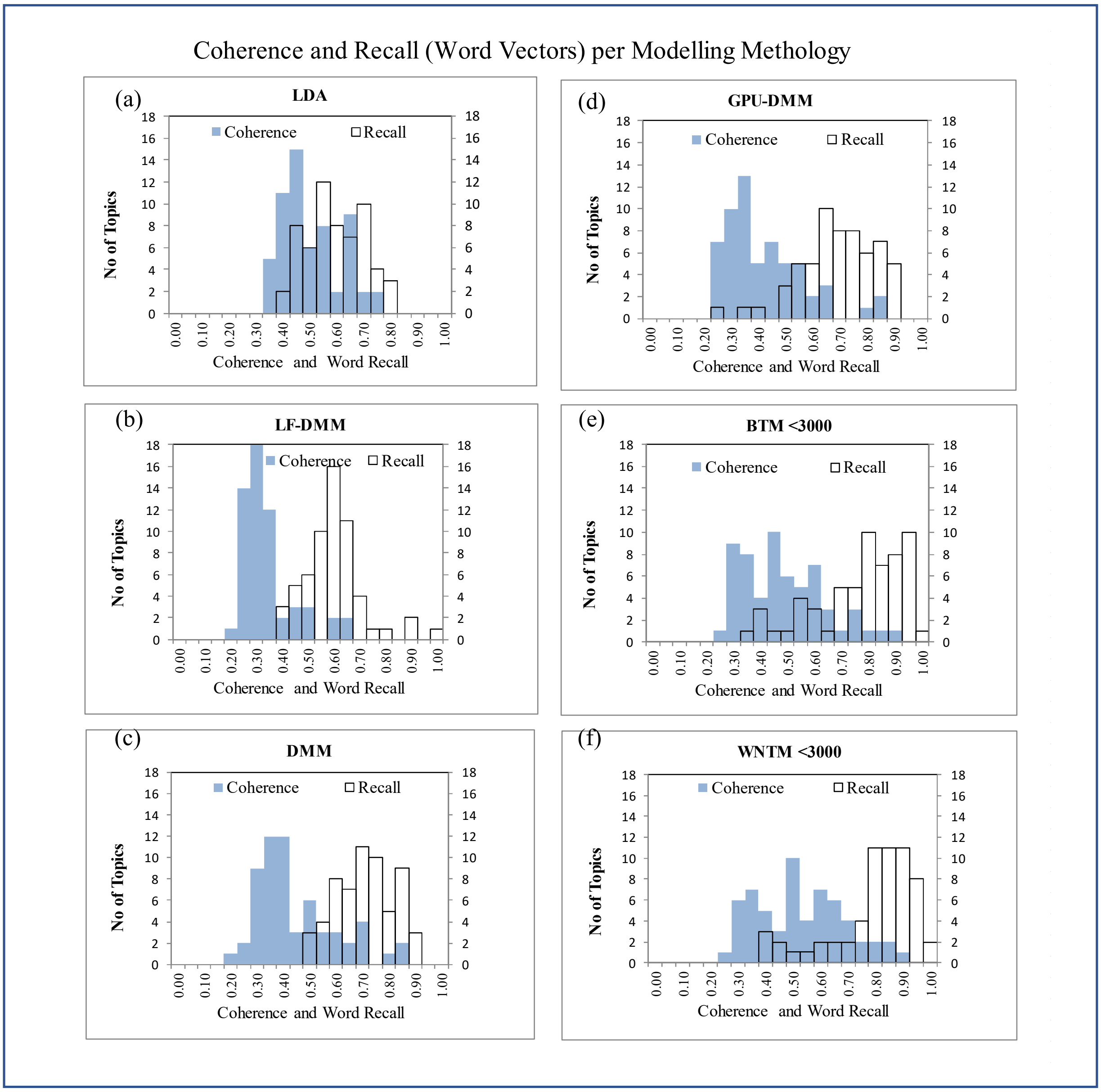

Figure 8: Graphs of Coherence Word Vector Classification Recall Values per Modelling Methodology.

Classes with more than 3,000 document counts were randomly sampled and limited to 3,000 documents. Both modelling methods underwent multiple runs. The macro averaged classification recall score for WNTM increased from 0.70 to 0.77, while the macro averaged classification recall score for BTM remained 0.74. The higher macro averaged classification recall scores added further validity to the better overall performance of these two topic modelling methodologies, in comparison to the other STTM methodologies. Coherence score distributions and classification recall distributions for topics can be seen in Figure 8 above.

5.1.5 Chi-squared Test

For BTM and WNTM chi-squared test of independence was performed on the topics. Scikit-learn was used to implement the chi-squared tests. The chi-squared test was applied to find the words most correlated with each topic. The word correlation allowed for both for further exploration into a given topic's content and to examine whether the most correlated words were consistent with the theme of the topic – adding to the validity. The top 100 words per topic were examined. For both methods, chi-squared results returned additional words that were consistent with topic themes that had higher coherence scores. The chi-squared output also gave further insight into the nature of a given topic. Chi-squared testing on the WNTM Nobel Prize topic helped reveal that Nobel Laureates, like Malala and Obama, were also brought up in the conversations. For the WNTM Montreal climate strike topic, chi-squared testing quickly helped reveal the support for and the organisation of climate marches in other cities, like Dhaka, Melbourne, Auckland, and Halifax.

6 Discussion

In the discussion section, performance metrics are summarised, and the topic output is assessed.

6.1 Discussion on Topic Modelling performance

Utilising the more appropriate C_v metric, the topics of WNTM and BTM were shown to be as coherent as topics from LDA with no significant differences in post hoc testing. These results were different from other studies, such as [21] where WNTM was found to be significantly better than LDA and BTM, and [29] where BTM was found to outperform LDA. The discussed papers did not use reliable coherence metrics. The window size for coherence evaluation context and use of an external corpus for coherence evaluation were not considered, while, using an outdated external corpus gives unreliable results [35].

To date, there has been no study making use of word vectors in SVC classification to compare the implemented algorithms. WNTM and BTM outperform the other topic modelling algorithms in the average classification recall by at least 11 per cent and 7 per cent respectively. This better performance was possibly due to the lack of consideration for sparsity within LDA [33]. Likewise, the independent studies of [21], [29], found that BTM and WNTM outperformed LDA, albeit with topic proportions as features and Hashtag labelled data – such use of a Hashtag as a label can lead to misleading results.

These studies did not use a combination of cross-reference metrics of combined coherence testing and classification recall to filter out topics with a lower degree of confidence. It was found that approximately sixty per cent of individual documents in the corpus were excluded, mainly because the topics with high document counts did not always have an associated coherent topic. The approach taken in this research project has shown that such a quality control mechanism significantly improves the confidence in the identified topics, where they meet the combined set criteria of coherence score and classification recall. The coherence words identified also provide relevant context to the top-twenty words in each of the identified topics.

Further analysis and topic modelling was undertaken on these low confidence Tweets. WNTM was rerun on the 350,000 Tweets using the same parameters, and class balancing as before, except that k was changed to 50. The same combined classification recall and coherence scores were used to identify valid topics. Approximately 10% of the original corpus was modelled into valid topics. The results showed that the rerun could yield valid topics; the majority have subject matter links to the previously discovered topics from the first run. There were a few new topics or subtly different topics. The topic modelling algorithm can be rerun to extract multiple sets of results, but it was expected that further returns would be diminished.

6.2 Discussion on the Topics Generated

There were 570,000 Tweets evaluated. The corpus dictionary was extensive at 29,000 words. Between 560 and 700 unique words were identified out of a potential 1200 in the top-twenty topic defining words for each topic modelling method (except for LF-DMM, which only had 290 unique words). Most of the words that rank in multiple topics top-twenty words have a high word count in the corpus – typically above 25,000 words each. Examples of such words include: *like, one, know, get* and *think*. Following the initial examination of the corpus before the processing phase, it had been previously concluded that these ancillary words added to the discussion on climate change, as they helped maintain the correct interpretation of the short texts.

The topics identified by the various models showed some common themes. These included topics related to Greta's activities, such as the UN Climate Action Speech, a rallying cry for Friday Strikes, her recent Nobel Prize nomination and her Atlantic voyage. Many Tweets related to different aspects of Greta's speech, which were discussed in multiple topics. One topic focused on the scathing tone of Greta's speech, while a few topics focused on the content of the speech. One such topic related to content of Greta's speech focuses on people in power and the older generation failing to act.

There are several negatively leaning, often sarcastic topics. Examples of disparaging topics include Greta being a divine figure, the abuse/misuse by her parents, that she should go back to school, getting funding from George Soros, some political commentators making irrelevant and unfavourable comparisons to people such as Hitler and Stephen King's *The Children of the Corn*, and comparisons to the right-wing Covington High School students. Several topics concerned discussions about Greta's mental state, such as her diagnosis with the Asperger's syndrome, autism, and mutism; as well as using these conditions as insults to detract from the sensible discussion that should have been taking place.

Although one would expect a significant discussion on the merits of the concept of climate change, contrarian detractors on Twitter often put more effort into personal attacks on Greta herself, rather than putting forward counter-arguments to her clear message and call for action. These personal attacks are in line with those that Al Gore experienced following the release of his 2004 documentary *An Inconvenient Truth* [1].

The time series plots in Figure 9 and Figure 10 below show that the volume of Twitter responses on all topics peaked just after 14h00 on 23 September 2019 (Coordinated Universal Time or UTC) with a progressive decay in traffic over five days. The daily cyclicity of the Twitter volume was also visible and represented the effect of multiple time zones of the discussion and the dominance that the USA users have on the platform. The proportion between the individual topics for the LDA model was very constant in Figure 9, with each topic apparently tending to respond proportionally to changes in the total daily traffic volume.

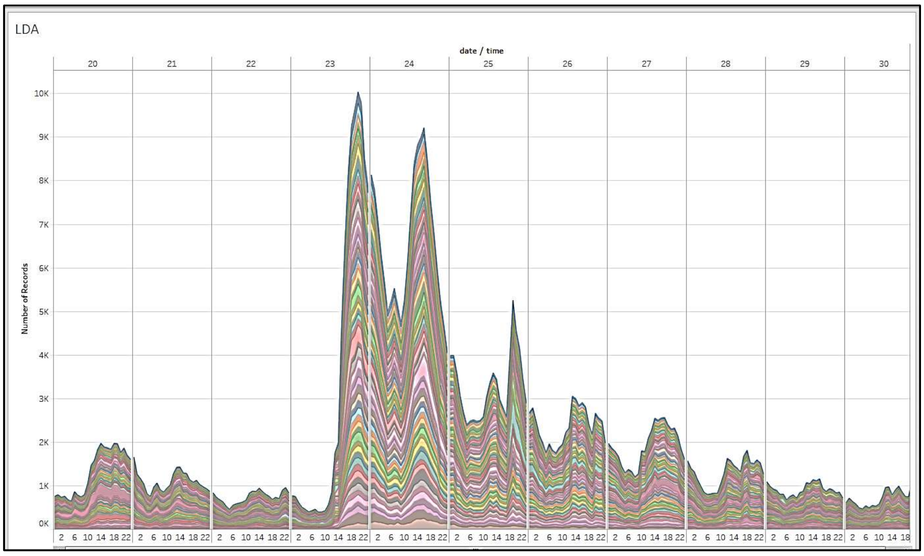

Figure 9: Time series plot of the topic modelled by LDA.

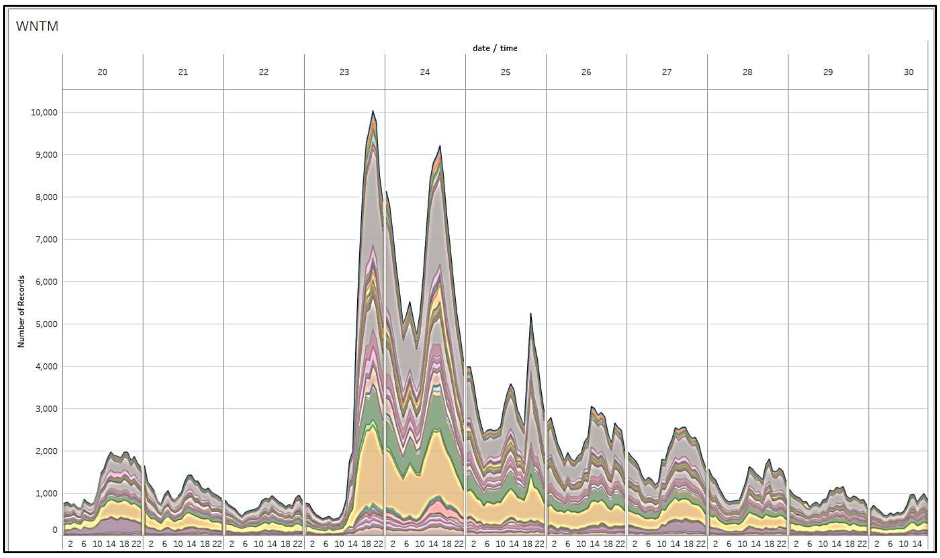

Figure 10: Time series plot of the topic modelled by WNTM.

WNTM however, shows significant variability in proportions of specific topics in Figure 10, where there are three numerically large topics. Figure 10 also highlights that the proportion of the more dominant topics varies with time in the WNTM topic modelling, to a greater extent than that of LDA.

In contrast to the backlash, there were many Tweets of support for Greta and Tweets attacking right-wing commentators. Many of these Tweets even formed a robust response, which stood out significantly in the time series plots on 25 September for both WNTM and BTM. The spike in the number of Tweets on the 25 September was most likely in response to the puzzling backlash of the previous few days. The build-up of Tweets for the topic can be seen for BTM and WNTM in Figure 11 below. Calls for Friday Climate Strikes with definitive peaks on each of the two Fridays helps validate the discriminative performance of LDA, BTM and WNTM. LDA, however, failed to accurately define the hourly frequency of the dominant congratulatory topic about Greta's activism.

Other topics discussed about Greta include several technical topics related to climate change, such as the footprint of biggest polluter countries, concern for the natural world and the environment, her re-iteration of scientific evidence, discussion about making a difference, discussion of climate change being a human-made issue, and the topic suggesting ways to lower one's carbon footprint. Some topics only mention Greta in passing. An example of such topic was a topic entailing Joe Biden's corruption scandal in Ukraine and the link to Donald Trump's Impeachment.

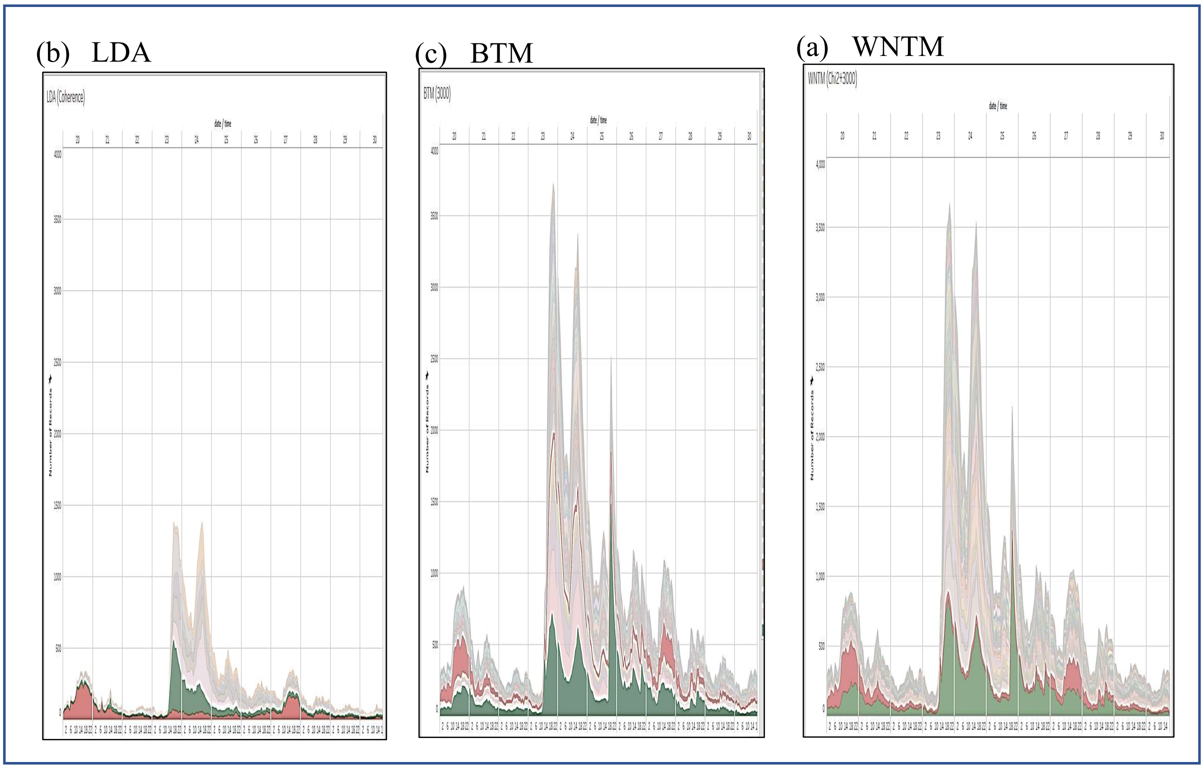

Figure 11: Time series plot of two main topics for LDA, BTM and WNTM Scale 4000 Tweets, Congratulatory (green) topic and Friday Strike (red) topic.

As stated in the literature review, a growing sense of polarisation was found by [8]. For the current study, there was a clear sense in polarisation upon examination of many of the topics. Some topics have a consistent theme, but have lots of discourse, while other topics are very partisan. There was a strong tendency for distractors of climate change to use abusive and degrading terminology to discredit the concept of climate change, that it was a hoax and that parents are misusing the children to further their own aims.

Although not treated as bigrams in the corpus, the words '*climate'* and '*change*' are of a much higher frequency than the words '*globa*l' and '*warming*', which was in line with observations by [8], showing that language and terminology related to a subject changes with time. This change was also politically driven by the US, as the terminology was coined by

Frank Lunz in 2002[19], to deflect attention from the real effect of *global warming*, with *climate change* appearing to be a more natural phenomenon.

7 Conclusion and Future Work

The project entailed two focus areas covered by the Research Question as to whether the short text topic modelling techniques DMM, LF-DMM, GPU-DMM, BTM, and WNTM could outperform LDA to reflect the properties of the document corpus on Twitter appropriately. The secondary Sub-Research Question focused on the common themes of discussion around Greta Thunberg concerning her address to the United Nations Climate Action Summit on the 23 September 2019. The short text topic modelling algorithms were run utilising the 570,000 Tweets collected over a ten day period associated with the UN Climate Action Summit. The collected Tweets contained the key search term '*greta*' to evaluate the impact Greta Thunberg had on the discussion around climate change. Many valid topics were generated. The short text topic modelling methodologies that showed the most valid topic modelling results were the non-generative methods of BTM and WNTM, followed by the GPU-DMM and DMM. The standard topic modelling methodology LDA, typically used for long documents, showed deceptive results. Although LDA produced good coherence scores, it was shown to be unreliable upon examination of the topic proportions and the word feature classification output. Time series plots of the 'Congratulatory' and 'Friday Strike' topics of the WNTM and BTM outputs showed a marked consistency in the modelling of the similar topics in relation to time, adding credibility to the effectiveness of the two topic modelling techniques. There were clear topical themes of support for Greta Thunberg, while many topics were showing a clear objection to Greta's activism, often having an unfavourable bias and of a personal nature. Both the primary and secondary research questions have been answered. From a stakeholder perspective, the understanding of the opposing views and the motivation of coordinated negative publicity campaigns can markedly improve the reach of a grassroots campaign, particularly for climate change advocacy. Additionally, the future use of WNTM will allow stakeholders, such as marketing firms and community focus groups, to monitor discussion on important topics without having the time or the money to spend on traditional polling campaigns. Future work could focus on finding an optimal topic number, and the systematic re-evaluation of Tweets belonging to topics with low coherence and classification recall scores. Topics from the generated data centring on the Climate Action Summit (September 2019) could also be compared and contrasted to topics generated from the Cop25[20] meeting in December 2019.

8 Acknowledgements

The author would like to thank Dr Catherine Mulwa for her guidance and direction throughout the project.

---

[19] http://aireform.com/resources/archive-2002-memorandum-to-bush-white-house-by-gop-consultant-frank-luntz-17p/
[20] https://unfccc.int/cop25

References

[1] C. Boussalis and T. G. Coan, 'Text-mining the signals of climate change doubt', *Global Environmental Change*, vol. 36, pp. 89–100, Jan. 2016, doi: 10.1016/j.gloenvcha.2015.12.001.

[2] J. Farrell, 'Corporate funding and ideological polarization about climate change', *Proc Natl Acad Sci USA*, vol. 113, no. 1, pp. 92–97, Jan. 2016, doi: 10.1073/pnas.1509433112.

[3] S. C. Moser, 'Communicating climate change: history, challenges, process and future directions: Communicating climate change', *WIREs Clim Change*, vol. 1, no. 1, pp. 31–53, Jan. 2010, doi: 10.1002/wcc.11.

[4] L. Collins and B. Nerlich, 'Examining User Comments for Deliberative Democracy: A Corpus-driven Analysis of the Climate Change Debate Online', *Environmental Communication*, vol. 9, no. 2, pp. 189–207, Apr. 2015, doi: 10.1080/17524032.2014.981560.

[5] E. M. Cody, A. J. Reagan, L. Mitchell, P. S. Dodds, and C. M. Danforth, 'Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll', *PLoS ONE*, vol. 10, no. 8, p. e0136092, Aug. 2015, doi: 10.1371/journal.pone.0136092.

[6] X. An, A. Ganguly, Y. Fang, S. Scyphers, A. Hunter, and J. Dy, 'TrackingClimateChangeOnTwitterKDD_Twitter_2014.pdf', presented at the KDD Workshop on Data Science for Social Good, New York City, New York, USA, Aug. 2014.

[7] A. Reyes-Menendez, J. Saura, and C. Alvarez-Alonso, 'Understanding #WorldEnvironmentDay User Opinions in Twitter: A Topic-Based Sentiment Analysis Approach', *IJERPH*, vol. 15, no. 11, p. 2537, Nov. 2018, doi: 10.3390/ijerph15112537.

[8] M. Lineman, Y. Do, J. Y. Kim, and G.-J. Joo, 'Talking about Climate Change and Global Warming', *PLoS ONE*, vol. 10, no. 9, p. e0138996, Sep. 2015, doi: 10.1371/journal.pone.0138996.

[9] H. T. P. Williams, J. R. McMurray, T. Kurz, and F. Hugo Lambert, 'Network analysis reveals open forums and echo chambers in social media discussions of climate change', *Global Environmental Change*, vol. 32, pp. 126–138, May 2015, doi: 10.1016/j.gloenvcha.2015.03.006.

[10] W. Pearce, K. Holmberg, I. Hellsten, and B. Nerlich, 'Climate Change on Twitter: Topics, Communities and Conversations about the 2013 IPCC Working Group 1 Report', *PLoS ONE*, vol. 9, no. 4, p. e94785, Apr. 2014, doi: 10.1371/journal.pone.0094785.

[11] E. C. Leas *et al.*, 'Big Data Sensors of Organic Advocacy: The Case of Leonardo DiCaprio and Climate Change', *PLoS ONE*, vol. 11, no. 8, p. e0159885, Aug. 2016, doi: 10.1371/journal.pone.0159885.

[12] A. P. Kirilenko and S. O. Stepchenkova, 'Public microblogging on climate change: One year of Twitter worldwide', *Global Environmental Change*, vol. 26, pp. 171–182, May 2014, doi: 10.1016/j.gloenvcha.2014.02.008.

[13] G. A. Veltri and D. Atanasova, 'Climate change on Twitter: Content, media ecology and information sharing behaviour', *Public Underst Sci*, vol. 26, no. 6, pp. 721–737, Aug. 2017, doi: 10.1177/0963662515613702.

[14] J. Jung, P. Petkanic, D. Nan, and J. H. Kim, 'When a Girl Awakened the World: A User and Social Message Analysis of Greta Thunberg', *Sustainability*, vol. 12, no. 7, p. 2707, Mar. 2020, doi: 10.3390/su12072707.

[15] D. M. Blei, 'Latent Dirichlet Allocation', *Journal of Machine Learning Research*, vol. 3, p. 30, 2003.

[16] J. Boyd-Graber, Y. Hu, and D. Mimno, *Applications of Topic Models*. Foundations and Trends in Information Retrieval, 2017.

[17] W. M. Darling, 'A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling', p. 10, Dec. 2011.

[18] S. Stier, A. Bleier, H. Lietz, and M. Strohmaier, 'Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter', *Political Communication*, vol. 35, no. 1, pp. 50–74, Jan. 2018, doi: 10.1080/10584609.2017.1334728.

[19] D. Elgesem, L. Steskal, and N. Diakopoulos, 'Structure and Content of the Discourse on Climate Change in the Blogosphere: The Big Picture', *Environmental Communication*, vol. 9, no. 2, pp. 169–188, Apr. 2015, doi: 10.1080/17524032.2014.983536.
[20] L. Hobson, H. Hapke, and C. Howard, *Natural Language Processing in Action_ Understanding, analyzing, and generating text with Python.pdf*. Manning Publications, 2019.
[21] Y. Zuo, J. Zhao, and K. Xu, 'Word network topic model: a simple but general solution for short and imbalanced texts', *Knowl Inf Syst*, vol. 48, no. 2, pp. 379–398, Aug. 2016, doi: 10.1007/s10115-015-0882-z.
[22] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, 'Probabilistic author-topic models for information discovery', in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, Seattle, WA, USA, 2004, p. 306, doi: 10.1145/1014052.1014087.
[23] D. Alvarez-Melis and M. Saveski, 'Topic Modeling in Twitter: Aggregating Tweets by Conversations', in *Tenth International AAAI Conference on Web and Social Media*, Mar. 2016, p. 4.
[24] E. Jonsson and J. Stolee, 'An Evaluation of Topic Modelling Techniques for Twitter', p. 11.
[25] J. Yin and J. Wang, 'A dirichlet multinomial mixture model-based approach for short text clustering', in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, New York, New York, USA, 2014, pp. 233–242, doi: 10.1145/2623330.2623715.
[26] Q. Jipeng, Q. Zhenyu, L. Yun, Y. Yunhao, and W. Xindong, 'Short Text Topic Modeling Techniques, Applications, and Performance: A Survey', *IEEE Transactions on Knowledge and Data Engineering*, May 2020, doi: 10.1109/TKDE.2020.2992485.
[27] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, 'Topic Modeling for Short Texts with Auxiliary Word Embeddings', in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16*, Pisa, Italy, 2016, pp. 165–174, doi: 10.1145/2911451.2911499.
[28] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, 'Improving Topic Models with Latent Feature Word Representations', *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 299–313, Dec. 2015, doi: 10.1162/tacl_a_00140.
[29] X. Yan, J. Guo, Y. Lan, and X. Cheng, 'A Biterm Topic Model for Short Texts', *Association for Computing Machinery*, vol. WWW '13: Proceedings of the 22nd international conference on World Wide Web, p. 11, May 2013.
[30] M. Röder, A. Both, and A. Hinneburg, 'Exploring the Space of Topic Coherence Measures', in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, Shanghai, China, 2015, pp. 399–408, doi: 10.1145/2684822.2685324.
[31] F. Colas and P. Brazdil, 'Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks', in *Artificial Intelligence in Theory and Practice*, vol. 217, M. Bramer, Ed. Springer US, 2006, pp. 169–178.
[32] T. Joachims, 'A statistical learning model of text classification for support vector machines', presented at the SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, Sep. 2001.
[33] V. Rus, N. Niraula, and R. Banjade, 'Similarity Measures Based on Latent Dirichlet Allocation', in *Computational Linguistics and Intelligent Text Processing*, vol. 7816, A. Gelbukh, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 459–470.
[34] A.-S. Pietsch and S. Lessmann, 'Topic modeling for analyzing open-ended survey responses', *Journal of Business Analytics*, vol. 1, no. 2, pp. 93–116, Jul. 2018, doi: 10.1080/2573234X.2019.1590131.
[35] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, 'Optimizing Semantic Coherence in Topic Models', *Association for computational linguistics*, p. 11, 2011.