# Geographical Sentiment Distribution For Coronavirus

Adam Sadiq, Aaron Wray, Anan Shekher Srivastava, Ashwinder Khurana,
Dominic Rawlins, Samuel Martin
Emails : {as16165, aw16997, wn19994, ak16625, dr16639, sm16616}@bristol.ac.uk
Department of Computer Science,
University of Bristol
*GitHub Repository: Coronavirus-Sentiment*

***Abstract*— Twitter has 330 million monthly active users, contributing to the 500 million tweets sent per day [1]. This increasing availability of large volume social media data has motivated research into applying sentiment analysis and natural language processing to understand how people react to different topics. This paper looks at how people across the globe have reacted to the Corona-virus pandemic. We aim to analyse the level of panic by analysing user tweets. Further, we also discuss how (the level of) panic has changed over time, across different regions.**

## I. Introduction

Twitter has 330 million monthly active users, contributing to the 500 million tweets sent per day [1]. This increasing availability of large volume social media data has motivated research into applying sentiment analysis and natural language processing. On top of the volume of users, the landscape of the users makes the data from the social media site particularly attractive to analyse because users are spread all across the globe, which is well suited for the geographic distribution of sentiment.

COVID-19 has had no shortage of discourse and controversy, and thus evolving sentiment, particularly concerning politics, general health and welfare. This paper focuses on the geographic distribution of panic as different countries react to the rapidly unfolding effects of the outbreak of COVID-19. The panic sentiment is of interest as there has been a multitude of attitudes toward the outbreak, which have inadvertently affected how different governments and populations of countries have reacted to it or vice versa.

This paper aims to model the panic distribution geographically utilising twitter data with state of the art machine learning and sentiment analysis techniques to narrate the panic levels, particularly around the day of the announcement from the World Health Organisation (WHO), classifying COVID-19 as a pandemic.

## II. Data preprocessing

### A. Data Collection

The official Twitter API, created by Twitter for developers, allows users to search through historical tweets. The official API requires authentication and has several tiers of usage: standard, premium and enterprise. Both premium and enterprise tiers are paid-for services. Therefore, neither tiers were considered when collecting data for this project. Additionally, the standard tier only offers searches against a sample (with a hard limit on the number of tweets) of recent Tweets published in the past seven days. Thus, the Twitter API was considered an insufficient source of information for this project.

Alternatively, there are libraries written in Python that do not use Twitter's API to retrieve tweets and require no authentication. For example, the Twint and twitterscraper libraries both use HTML GET requests to retrieve a page of tweets containing a chosen keyword from a URL [2], [3]. Once the libraries have finished scraping all tweets from one URL, the page number is then incremented to collect more tweets until there are no pages remaining for a specific keyword. Both libraries impose no restrictions on retrieving tweets and allow a user to select a date range from when to collect tweets. The Twint library was preferred as it offers greater usability, allowing a user to select the required parameters to retrieve from tweets. There are many parameters to choose from, but the ones required for this project are: location data, tweet text and username.

The Twint library provides many benefits when collecting tweets and their required metadata. As previously mentioned, the library uses HTML get requests to retrieve a page of tweets. The library also uses HTML get requests to retrieve a user's profile in order to obtain the location data for every tweet. This implementation is inefficient and leads to approximately 2000 tweets being collected per hour. For this reason, the decision was made to use Amazon's Elastic Compute Cloud service to parallelise data collection using up to 32 virtual machines at a time.

There are a large number of tweets available every day, and for the scope of this project, it was not feasible to collect all tweets containing the keywords ”coronavirus” and ”covid-19” in the year of 2020. Therefore, the decision was made to collect tweets on two specific dates to capture the shifts in sentiment in different locations. On March $11^{th}$ 2020, the World Health Organisation (WHO) declared the coronavirus as a pandemic [4]. The following day (March $12^{th}$ 2020) and the day prior (March $10^{th}$ 2020) were selected to collect tweets and accurately gauge the sentiment of the global reaction to this announcement.

### *B. Data Cleaning*

Feature selection and pre-processing text data are essential steps in the data pipeline, as they transform the textual data into a form the machine learning models understand. Both can have a tremendous impact on the success of these models, especially in natural language processing (NLP). This is mostly because text data in nature is highly unstructured, whilst machines need structured and numerical data. Tweets are especially messy since they are typically short and informal. The informality problem ranges from the writing style of different users to the existence of slang, emoticons, abbreviations and bad use of punctuation. This section's scope lies in the rigorous pre-processing and feature selection techniques used and the reasoning for using each.

Tweets often co-occur with a lot of extraneous data such as invalid characters, punctuation and blocks of white spaces. Therefore, this extraneous data is removed as it is uninformative to the machine learning models; it merely adds noise and reducing noise in the text should help to improve the clustering performance and speed.

The next step is tokenisation which is considered an inevitable part of pre-processing text data: it is used to split longer strings of text into smaller pieces called tokens. At the same time, each token is checked for whether it is a stop word and is thrown away if so. After that, each token is lemmatised meaning that it will be returned to its root word if possible. Stop words, such as ”the”, ”a” and ”is”, are removed because they do not carry much discriminative content; they do not help in determining the topic of a tweet for the models used. Thus, it helps to reduce the raw input space by reducing the number of features. Lemmatization also helps to reduce the input space by mapping different word forms to their common representation. For example, 'runs', 'ran' and 'running' will all be returned to 'run'. Reducing the input space allows the models to focus on the features that provide the most useful information, which should improve performance. Tokenisation, stop word removal, and lemmatisation was all implemented using the NLTK python library.

### *C. Feature Extraction*

Text clustering algorithms such as GMM and LDA (used in this project) cannot work with text in a raw form, meaning that we had to convert the input text into numeric form. The type of numerical text representation can be essential to the performance of the machine learning models. We opted to use the most straightforward text representation, the bag-of-words (BoW) model, which is commonly used in many text classification and clustering tasks.

The bag-of-words model is a vector space representation (a multidimensional document-term matrix) which creates a vocabulary consisting of all the unique words occurring in the corpus of documents (tweets in our case). This vocabulary corresponds to the number of dimensions (columns/ features) in the matrix, and the tweets in the corpus correspond to the rows where each row (tweet) is represented as a numerical feature vector. The value for each feature for a given tweet is the number of times the feature appears in that tweet, which is called term frequency (TF) [5]. This is shown by the following equation:

$$tf(t,d) = f_{t,d} \tag{1}$$

Where $t$ is a term, $d$ is a document (or a tweet) and $f_{t,d}$ the number of times that term $t$ occurs in document d.

It is common to normalise term frequencies based on their presence in the whole corpus of documents, in order to avoid bias. To do this, inverse document frequency or IDF can be used which diminishes the weight of terms that occur very frequently. This can be shown by the following equation:

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{2}$$

Where:
$N$ denotes the total number of corpus N = $|D|$, $|\{d \in D : t \in d\}|$ denotes the number of documents where the term $t$ appears. [6]

IDF can then be applied to TF to produce term frequency-inverse document frequency (TF-IDF):

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{3}$$

The TF-IDF value increases proportionally to the number of times a word appears in a tweet and is offset by the number of tweets in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general [6].

Using the BOW model with term frequencies was sufficient as input to the GMM and preferred to TF-IDF. This is because TF-IDF produces a lower value for a term appearing the same number of times in a tweet with more words. This will produce a lower value for panic tweets in longer tweets; it makes no sense that a longer tweet has a higher panic value. However, for LDA, this is unimportant, which makes TF-IDF very useful to this model.

We chose to extend the BOW model to an n-gram BOW model, which represents a text document as an unordered collection of its n-grams [7]. We did this because it served to be more informative than bag-of-words as it captures more context around each word. For example, bigrams such as ’coronavirus fear’ and trigrams such as ’crisis del coronavirus’ can help significantly in classifying tweets as panic or not panic. For our case, we found that an n-gram value of (1,3) worked best.

The bag-of-words model, TF, TF-IDF and n-grams were all implemented using sklearn’s python implementation called CountVectorizer and TfidfVectorizer. Through this, we were also able to remove words that rarely occurred (in 3 or fewer documents/ tweets) using the min-df parameter. Since these words are rare, the association between them and other words is dominated by noise. Hence, they contribute little to nothing during model training.

## III. Gaussian Mixture Model Construction

### A. Gaussian Mixture Models

Gaussian Mixture Models (GMMs) are an unsupervised machine learning approach to clustering. GMMs provide a soft clustering method, calculating the probability an individual data point being associated with a given cluster, where the probability of the data point, X, is a vector that represents the probability of the data point belonging to different clusters. Therefore, a point can partially belong to multiple clusters, giving a sense of uncertainty. Uncertainty can inform decision making. In a GMM it is assumed that there are $n$ Gaussian distributions with each of these $n$ distributions representing a cluster. The formal definition for a GMM can be seen as follows:

$$p(X) = \sum_{k=1}^{K} p(\mathbf{X}|k) = \sum_{k=1}^{K} \mathcal{N}(\mathbf{X}|\mu_k, \Sigma_k)p(k)[8] \tag{4}$$

where:

$X$, $k$, $\mu_k$, and $\Sigma$ are the data, selected distribution, mean and covariances of the distribution $k$.

The Expectation Maximisation algorithm is then applied to the model to provide an optimal estimate of the unknown parameters ($\mu$, $\Sigma$, $\pi$) for each Gaussian distribution. Each data point then has a probability of being a part of each cluster. The process is as follows:

*a) Expectation Step::* The algorithm starts with random Gaussian parameters ($\theta$). It then computes $p(z_i = k|x_i, \theta)$

*b) Maximisation Step::* We want to maximise the likelihood that the parameters generated came from the distribution of data.

$$max(\prod_{i=1}^{N} p(x_i|\theta)) \tag{5}$$

We then update the Gaussian parameters ($\theta$) and repeat the EM steps until convergence.

### *B. GMMs to Predict Panic Globally*

The approach within this section was based off work done by Xu and Qiu in 2019 [9]. An unsupervised approach was appropriate for this task due to both the lack of labels for the data and the time investment it would have taken to label hundreds of thousands of tweets. The aim of the GMM was to classify tweets into two clusters: panic and non-panic tweets.

Teets from each source were grouped to implement this, and the bag of words applied to each group. For example, if the desired approach were to look at the geographic spread of panic sentiment, all the tweets from a particular country would be grouped together and a bag of words vector created for this collection of tweets. To minimise RAM usage, each tweet can be turned into a bag of words vector, with the bag of words vectors for each group summed up iteratively; this forewent the need to load all tweets into memory and apply the bag of words holistically. This approach yields a bag of words vector for each group. With each data item in a vector representing a dimension in the GMM model, a large bag of words will mean a high dimensionality of the GMM model. To reduce the dimensionality of the model and drastically reduce the computation needed, Principal Component Analysis (PCA) is used. PCA reduces the dimensionality of a dataset with a large number of related variables, whilst still retaining most of the variation present. This dramatically reduces the computational exercise for other models to learn from the dataset.

This can be done through computing the mean, covariance and eigenvectors from eigenvalues from the dataset. The highest eigenvectors will indicate those variables which provide most of the variation in the datasets and are thus principal components.

Let A be a square matrix, $v$ a vector and $\lambda$ a scalar that satisfies $Av = \lambda v$. The calculation of the eigenvectors and eigenvalues are calculated fulfilling this condition, where the eigenvalues of A are roots of the characteristic equation:

$$det(a - \lambda I) = 0 \tag{6}$$

In the analysis for this method, the data was reduced to two principal components. A scalar was used before this on the vectors to maximise PCA effectiveness.

Each groups vector then becomes a data item from which the GMM can learn the two clusters for panic and non-panic. Training this model will thus produce two Gaussian distributions. Artificial vectors can, at this point, be used to determine which cluster produced is which; two bag of words are created, one for the positive words, one for the negative words. The model can then be retrained until these two artificial vectors represent two different classes, implying the model has successfully differentiated the two sentiments. From this, each groups probability of being part of the panic cluster represents the level of panic their tweets contain.

### *C. Verifying the GMM model*

To verify the GMM model data, 250 tweets were taken from each day between the 11th of January and the 5th of April inclusive. These tweets were then modelled as previously described with the panic being plotted for each of these days in Fig.

1. This was plotted alongside the closing prices each day for the FTSE 100.

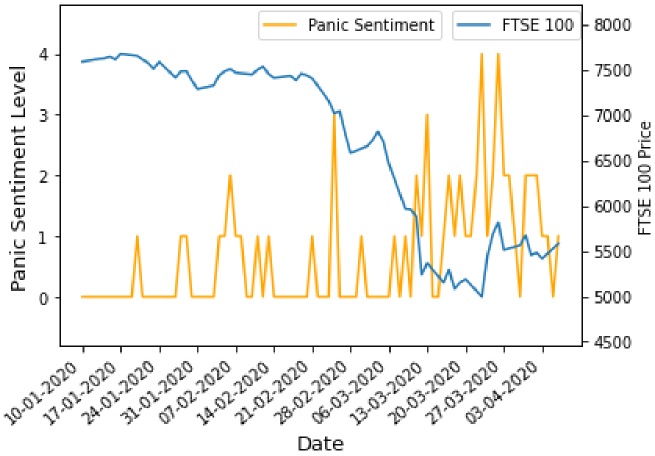

Fig. 1. A plot showing the panic sentiment over time

Pearson's Correlation Coefficient can be calculated using the equation:

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{7}$$

where $cov(X,Y)$ is the covariance of $X$ and $Y$, and $\sigma_X$ is the standard deviation of $X$. This can be used to calculate the correlation between two variables. The FTSE 100 and panic level has a Pearson Correlation Coefficient of -0.55(2dp) indicating there is a negative correlation between these variables. This indicates that sentiment of increasing panic about the coronavirus could have caused share prices to fall. The stock markets falling amid panic about coronavirus has been well documented, and this suggests that the model results are sufficiently accurate [10].

### D. *Key Challenges with GMM approach*

One key challenge presented by this approach was the certainty of the Gaussian. The GMM was very confident in the predictions for either class, meaning the model became a binary classifier. In reality, panic is better modelled as a scale. To combat this, and provide more nuanced results, each group's tweets were split into five groups. Each group was added to the model independently, with a group's panic being the proportion of these sub-groups that were classified as belonging to the panic cluster. In addition to this, the GMM's covariance matrices were the same to avoid overfitting and the model classing almost all data points to one cluster, losing information.

A key challenge to this approach was the lack of data for individual countries. Scraping over 1 million tweets yielded only 58,340 with location data. Of these 58,340 only 33,509 were English language tweets. The bag of words model requires English language tweets as it matches the English words provided. If the foreign language tweets were not filtered out, then the data would be skewed with potential panic tweets not appearing in the bag of words. Modelling globally presented a challenge with this data; the vast majority of tweets were from the U.S. with California alone providing over 3,300 of these tweets. This meant that many countries had a limited amount of tweets. In the end, 29 countries were modelled with each one having at least 40 tweets.

With a bag of words approach, the more tweets that were modelled, the higher the count in the bag of words was likely to be. As the number of tweets for each country in the dataset has no impact of the level of panic in the country, this was mitigated by making each bag of words proportional to the number of tweets supplied to create it. This meant if there were ten panic words in 100 tweets had the same value as 200 panic words in 2,000 tweets. This approach, however, made a panic word in a group with low tweet count extremely important. For example, a count of 1 use of the word 'scared' in 20 tweets gave a proportional value of 0.05 whereas five uses of the word 'scared' in 2000 tweets only produced a value of 0.0025. This gave rise to artificially large vectors for groups with a small tweet sample size. To mitigate this, a maximum value was applied to the proportional bag of words vector to make sure noise in the data at the lower end did not result in substantial values skewing the data and affecting the clustering algorithm.

### E. *GMMs to assess a current news story*

A conspiracy theory that claimed traction throughout the coronavirus pandemic was that the new 5G network was causing the virus [11]. This conspiracy theory has been proven false scientifically and so relied on a large amount of distrust in the establishment and misinformation by its supporters. President Donald Trump has oft stoked this distrust in his supporters as well as commonly giving misinformation [12], [13]. With an abundance of U.S. tweets, this model can be used to assess whether the level of support for Trump affects the likelihood of anti-science conspiracy tweets about coronavirus and that his anti-establishment rhetoric can influence a states sentiment.

The same approach was used as previously described; however, with a different bag of words: half the words pertaining to hoax-like rhetoric, the other half pertaining to science rhetoric. The level of hoax-like rhetoric was then compared to vote share Trump received in that state in the 2016 Federal elections. This is shown in Fig. 2

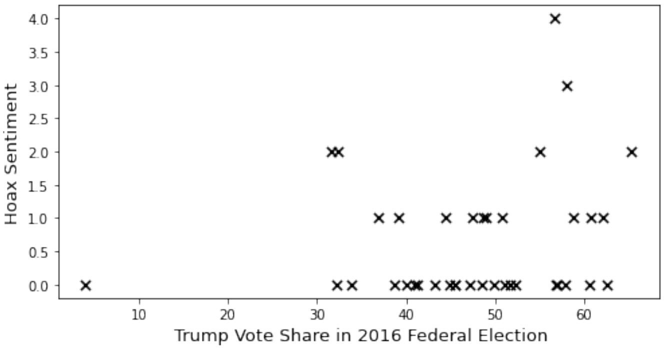

Fig. 2. A plot showing a State's twitter data hoax sentiment compared to the vote share Trump received in that state

The data suggests a slight positive correlation with a Pearson's Correlation Coefficient of 0.19 (2dp) and so suggests a link between a state's endorsement of President Trump and their sentiment on coronavirus.

There are some drawbacks to the GMM approach, which may mean the data is not more strongly correlated. Firstly, studies have shown that the average tweeter is left-of-centre and so far less likely to listen to a Republican president, such as Trump, and so this will mean the sentiment of tweets will be less varied [14].

Secondly, and more crucial from a data science perspective, is a downfall with the bag of words approach. The bag of words approach has no understanding of semantics; it merely counts the appearance of a word. This means that the tweets 'the coronavirus is a hoax' and 'the coronavirus is not a hoax' would both contribute to the hoax sentiment with this approach. This means that this approach has the possibility of being inaccurate in its results. It is also worth noting that a report such as 'conspiracy theorists claim 5G causes coronavirus' would also contribute to the hoax sentiment when it is clear that the tweet itself is not necessarily in favour of the opinion. A more sophisticated model that aims to capture the semantics of a sentence and word order better may generate more robust results.

## IV. Latent Dirichlet Allocation

### A. Introduction

In the overall analysis of the data, the usage of Gaussian Mixture Models as a standalone model does not provide the validity and the credibility required, so Topic Modelling was used as a means to supplement the results in Section III-A to provide a level of credence to the overall results. The goal behind the usage of Topic Modelling is to indicate that the overall sentiment and the topics being discussed prior and after the World Health Organisation announcement (classifying COVID-19 as a pandemic) are drastically different and more serious in their themes, which would indicate an overall increase in panic. This could be further supported by an attempt to identify the source of panic that are prevalent in each day.

### B. Technical Content

'Topic modelling', a statistical model to discover the abstract "topics" that are present in a corpus (a collection of documents). Latent Dirichlet Allocation (LDA) is the process by which the topics are discovered; the 'latent' part of the process alludes to the fact that the topics are discovered as part of the process and are not known prior. LDA is a 'Generative Process' and a form of unsupervised learning. From LDA's perspective, the tweets collected could be seen as being generated from an underlying complex process, and the LDA tries to model this via a synthetic process, whilst finding the parameters to the synthetic process that approximates the real documents - in this case, being tweets.

Fig. 3 outlines a graphical model for the LDA. Where:

$k$ denotes the number of topics a document belongs to (a fixed number),

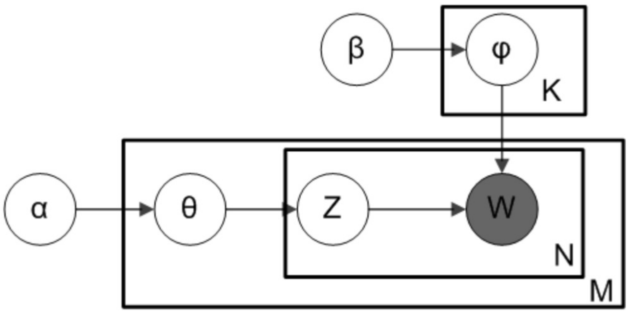

Fig. 3. Graphical Model for Latent Dirichlet Allocation

$V$ denotes the size of the vocabulary, $M$ denotes the number of documents, $N$ denotes the number of words in each document, $w$ denotes a word in a document - represented as a one-hot encoded vector of size $V$ (i.e. V  vocabulary size), $\mathbf{w}$ represents a document (i.e. vector of ”w” s) of N words, $D$ denotes corpus, a collection of M documents, $z$ denotes a topic from a set of $k$ topics. A topic is a distribution words. For example it might be, Animal = (0.3 Cats, 0.4 Dogs, 0 AI, 0.2 Loyal, 0.1 Evil) $\theta_{ij}$ represents the probability of the $ith$ document containing a word from the $jth$ topic, and $\beta_{ij}$ is a random matrix where $\beta_{(i,j)}$ represents the probability of $ith$ topic containing the $jth$ word [**?**].

Given that the problem at hand is the synthetic process of approximating the process of the data creation, it leads to the overall conditional probability below:

$$P(\theta_{1:M}, z_{1:M}, \beta_{1:k}|\mathcal{D}; \alpha_{1:M}, \eta_{1:k}) \tag{8}$$

The equation above produces an intractable posterior, so variational inference is used to approximate the true posterior with a probability distribution of a similar form by minimising the KL-Divergence to convert it to the optimisation problem below:

$$\gamma^*, \phi^*, \lambda^* = argmin_{(\gamma,\phi,\lambda)} D(q(\theta, z, \beta|\gamma, \phi, \lambda)||p(\theta, z, \beta|\mathcal{D}; \alpha, \eta)$$

### C. Method

LDA had been used to identify and visualise the range of topics before and after the WHO pandemic announcement. The motivation behind doing so was to highlight a shift in the themes being discussed on the social network by capturing the panic increasing after the announcement. In reference to Section II, the same process is used for lemmatising, stemming, tokenising and removing stop words. This will ensure that the words for each topic from the output of the LDA will be consistent and independent from tense or possession.

In the initial stages of LDA, a dictionary is built with a Bag of Words (BoW) and compiled so that the word count and TF-IDF can be calculated. To build this dictionary, two varying approaches were considered which provide differing results. The most common approach (which will be referred to as the ’Original’ approach in this section) is to build the dictionary using the data itself to allow the model to create the dictionary based on the content of the tweets. This approach will give a more holistic view of the topics in the dataset. The second approach that was taken was to use a customised and bespoke list of words for the model to ’look out’ for, where only these words will be counted throughout the tweets. Both of these approaches were conducted and compared.

The potential issue with building a customised BoW is analogous to confirmation bias; it potentially paints a narrative that may not be representative of the actual data, which may lead to a heavy data bias. In addition, building a finite list of words may not be exhaustive enough to cover the variety of ’topics’ and ’sentiments’ being discussed on the social media platform.

The original approach ran in parallel to mitigate these issues, and directly compared to the BoW approach. However, a potential issue with the original approach is the fact that words might be captured that is not directly related to the overall panic sentiment; they can be noted as spikes or trends at the specific time which could potentially skew the data, this is further explained in section V.

A TF-IDF model is applied on the BoW corpus created from the dictionary, and from this, the weighting based on importance is calculated for each word. With this information, it is now possible to instantiate and train the LDA model to cluster the documents into topics within the dataset.

Initially, the problems with the LDA was that the topics were not entirely distinguishable from because there were overlapping words that occupied relatively high weights in most of the topics, e.g. words like 'kill'; however, this was mitigated by adjusting the $\alpha$ parameter to a low value. This is because the $\alpha$ parameter denotes the document-topic distribution; since tweets are short documents by nature, they are unlikely to have a high distribution of topics within a single tweet. Additionally, the number of $passes$ in the training of the model was increased to 10.

The results of the LDA topic models are shown in the Model Evaluation in Section V.

## V. Model Evaluation

*1) Gaussian Mixture Models:* Fig. 4 shows the output of the panic levels from 1-5 throughout the country. Because of the limitation of the proportion of tweets per country, some could not be collected for every single country in the world.

The United States is the current epicentre of the pandemic with 1.43 million cases; as of the 14th of May, the U.S. is facing 3 million jobless claims weekly taking the total number above 36 million files for job loss [15]. Seeing the figure, it has been rated as the highest for panic level throughout the world. This is also seconded with the United Kingdom, which is the second-highest country with the most coronavirus related deaths and the third behind Russia in cases.

A manifestation of the limitation of the model can be seen in the country of Nigeria. With the model, it has received a panic level of 0. Although the population is approximately 200 million, the country seems to be relatively unconcerned by the pandemic; however, vast amounts of people in the country are still facing police brutality leading to deaths, for exercising the current freedom laws put in place [16], which can correlate to high panic.

*2) Latent Dirichlet Allocation:* One day before the WHO announcement of COVID-19 being a pandemic, the twitter topics from the LDA model can be seen in Fig. 6, 7 and 8. Amongst these figures, two figures show topics that appeared in the analysis of the model, which both seemingly relate to the cancellation and postponement of major events that were supposed to take place. The root words such as 'postpon', 'season', 'suspend' and 'cancel' are prominent words in the topics, which is suggestive of the fact that the greater source of panic was not down to the virus itself, but of how the virus would impact the entertainment industry.

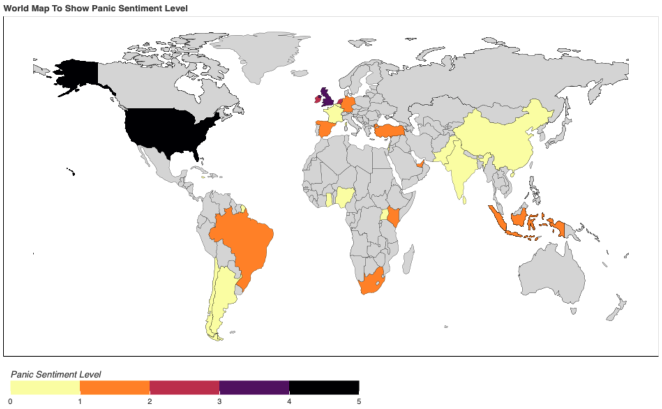

Fig. 4. World Map To Show Panic Sentiment Level.

In stark contrast, in the analysis of the tweets one day after the WHO announcement, the dominating topics revolve around the severity of the virus itself shown in Fig. 5 and 9. The virus itself being the dominant distribution of topics is suggestive that the seriousness of the pandemic had dawned amongst the countries that were predominantly tweeting in English, which would be namely: the U.S. and the U.K. This is seen in the profane language that makes up the word distribution in Fig. 9. Prior to the announcement, there were no profane words present in the model's results which can be interpreted to be indicative of growing frustration of the lack of preparation in some western countries, namely the U.K and the U.S.A [17], which is suggestive of panic increasing as a whole. There is still the underlying topic of event/sports cancellation, as seen in Fig. 5. This is likely due to to the closeness with the earlier topics from March $10^{th}$. However, with the increase in topics that are more serious in nature, it is evident that there is an increase in panic in comparison to the 10th of March.

Evidently, the topic distribution is not clear cut to draw definitive conclusions. However, the underlying themes can be seen that panic has been increased over time which fulfilled the motivation behind using the model, which was to indicate

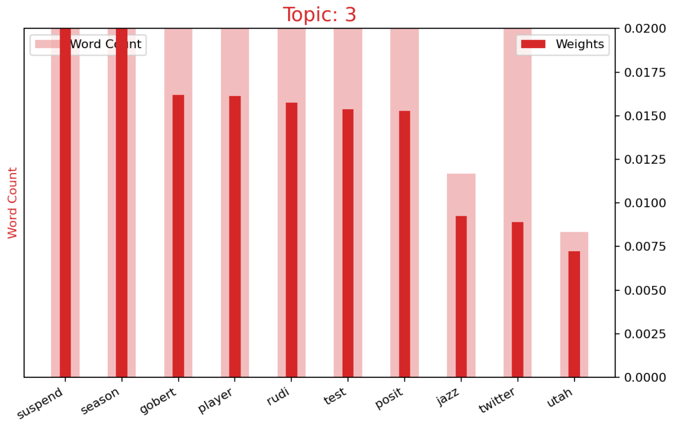

Fig. 5. Original approach showing Word Counts of Topic 3 one day after WHO announcement.

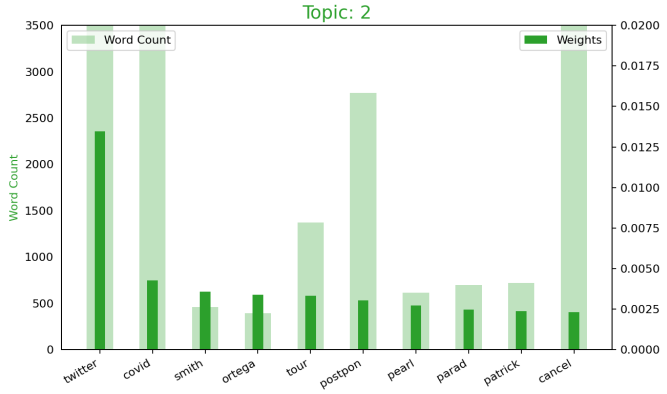

Fig. 6. Original approach showing Word Counts of Topic 2 one day before WHO announcement.

the source of panic and to support the GMM's conclusions.

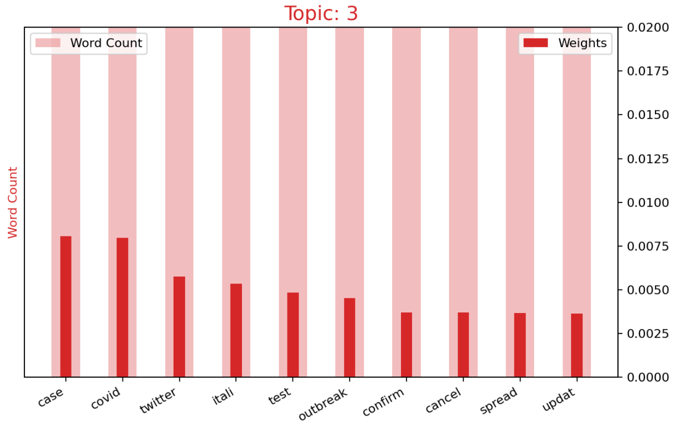

Fig. 7. Original approach showing Word Counts of Topic 3 one day before WHO announcement.

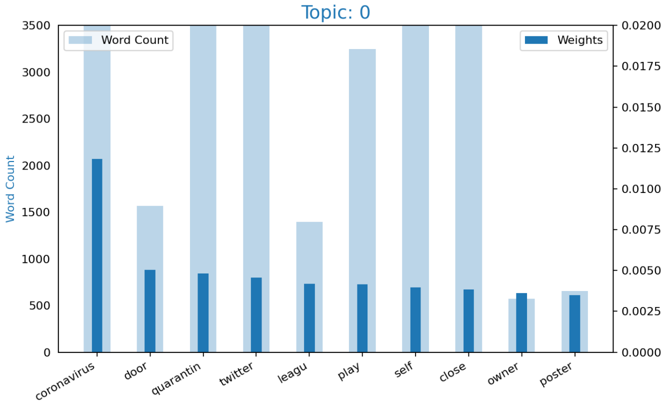

Fig. 8. Original approach showing Word Counts of Topic 0 one day before WHO announcement.

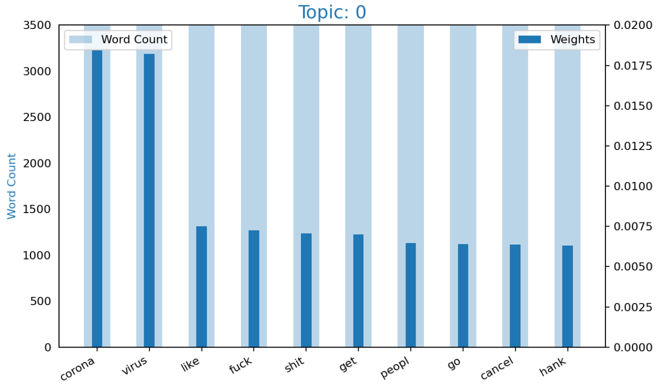

Fig. 9. Original approach showing Word Counts of Topic 0 one day after WHO announcement.

## VI. Future Work

The models created in this project focus specifically on the geographic distribution of sentiment after the WHO pandemic announcement. A continuation of this work could be to gauge the sentiment over a period of time. Additionally, other key events that have occurred since the beginning of the year 2020 that could be looked at more closely. For example, on the $23^{rd}$ January 2020, the Chinese Government put the city of Wuhan into lockdown. Wuhan is the city where the outbreak is thought to have originated. Also, on the $18^{th}$ March 2020 Canada and the U.S. close their border for non-essential travel. For the first time since the Canadian Confederation formed in 1867 [18].

Moreover, an investigation into the level of restrictions imposed (if any) on citizens across the world by their respective governments could be carried out. This could then be used to determine the sentiment towards the different restrictions that were put in place. The information obtained could be an informative and potentially useful way to gauge the level of ”acceptance” for the different measures taken to stop the spread of the coronavirus.

Since the beginning of the pandemic, there has been a large number of articles published in the mainstream media mentioning the term ”coronavirus” as shown in Fig. 10. The number of article mentions could be analysed against social media sentiment to establish any relationships between the sources.

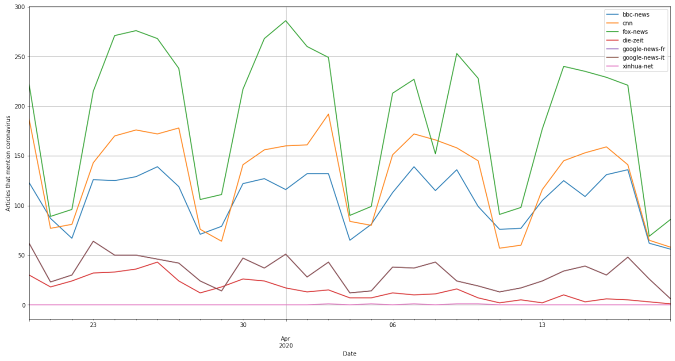

Fig. 10. Articles published in media concerning term “coronavirus”

The Bag of Words model has downfalls, as previously noted, in the ability to accurately calculate sentiment. For example, the sentence ’there is no panic’ would still be classified as panicked with a Bag of Words approach due to the presence of the ’word’ panic with no ability to pick up the effect of the word ’no’ on the sentiment. Future work would also aim to address this issue, potentially using such models as Doc2Vec to generate more nuanced vector representations of tweets to be used in the modelling stage [19].

## VII. Conclusion

The two main objectives of this paper were to indicate the level of panic caused by the Coronavirus pandemic, and how it changed over time, across the globe. We successfully create a machine learning model using the aforementioned techniques to achieve just that.

Whilst having several limitations, it is clear from these models that it is possible to view the sentiment towards a current topic, such as the COVID-19 pandemic, using data from a social media platform such as Twitter. It is also clear that the models do not have credence in isolation, but are better in conjunction with other datasets, e.g. financial and consumer data, to verify the sentiments captured in the models. Section VI outlines how improvements could be made, as well as additional supporting work if time permitted for this project. However, the current work shows that panic sentiment can be modelled, as well as its landscape, geographically, using social media.

## References

[1] Y. Lin, “10 twitter statistics every marketer should know in 2020 [infographic].” https://www.oberlo.com/blog/twitter-statistics, 2019.

[2] C. Zacharias, “Twint.” https://github.com/twintproject/twint, 2018.

[3] A. Taspinar, “Twitterscraper.” https://github.com/taspinar/twitterscraper, 2016.

[4] S. Boseley, “Who declares coronavirus pandemic,” 2020.

[5] http://barbra-coco.dyndns.org/yuri/Python/Machine+Learning+for+Text.pdf, title = Machine learning for text.

[6] “tf-idf.”

[7] “n-gram.” https://machinelearning.wtf/terms/bag-of-n-grams/.

[8] C. Ek, “Dirichlet processes,” 2018.

[9] L. Xu and J. Qiu, “Unsupervised multi-class sentiment classification approach,” *KNOWLEDGE ORGANIZATION*, vol. 46, pp. 15–32, 01 2019.

[10] P. Georgiadis, T. Stubbington, J. Rennison, E. Szalay, and S. Johnson, “How coronavirus tore through global markets in the first quarter.” https://www.ft.com/content/5f631cce-f75a-41d3-8f62-cff2fe83e90a, 2020.

[11] B. B. C. News, “Ofcom: Covid-19 5g theories are ’most common’ misinformation.” https://www.bbc.co.uk/news/technology-52370616, 2020.

[12] T. McCarthy, “’it will disappear’: the disinformation trump spread about the coronavirus timeline.” https://www.theguardian.com/us-news/2020/apr/14/trump-coronavirus-alerts-disinformation-timeline, 2020.

[13] J. E. Moreno, “Trump hits cnn and washington post reporters as ’fake news’ during briefing.” https://thehill.com/homenews/administration/494426-trump-hits-cnn-and-washington-post-report 2020.

[14] D. Freelon, “Tweeting left, right center: How users and attention are distributed across twitter,” *Knight Foundation*, 2019.

[15] L. A. Dominic Rushe and A. Holpuch, “36m americans now unemployed as another 3m file for benefits,” *The Guardian*, 2020.

[16] “Nigeria security forces have killed 18 people while enforcing coronavirus lockdown - when covid-19 has only claimed 12 lives in the country.” https://www.dailymail.co.uk/news/article-8225583/Nigeria-security-forces-killed-18-people-enforci html, 2020.

[17] S. Neville, “How poor planning left the uk without enough ppe — free to read.” https://www.ft.com/content/9680c20f-7b71-4f65-9bec-0e9554a8e0a7, 2020.

[18] C. Nardi, “covid-19 crisis canada u.s. border shut.” https://nationalpost.com/news/canada/covid-19-crisis-canada-u-s-border-shut-for-first-time-since-9-11-attacks, 2020.

[19] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” 2014.