The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-4/W1-2022
Free and Open Source Software for Geospatial (FOSS4G) 2022 – Academic Track, 22–28 August 2022, Florence, Italy

MAPPING THE CHATTER: SPATIAL METAPHORS FOR DYNAMIC TOPIC MODELLING OF SOCIAL MEDIA

L. Morandini[1,*], A. R. Mohammad[1], R. O. Sinnott[1]

[1] School of Computing and Information Systems, The University of Melbourne, Parkville VIC 3010, Melbourne, Australia - (luca.morandini, abdulrehman.mohammad, rsinnott)@unimelb.edu.au

Commission IV, WG IV/4

KEY WORDS: Dynamic Topic Modelling, Social Media, Spatialization, Twitter, Information Visualization, Model Validation.

ABSTRACT:

Topic modelling is a branch of Natural Language Processing (NLP) that deals with the discovery of conversation topics in a given document corpus. In social media, this translates into aggregating social media posts, e.g. tweets, into topics of conversation and observing how these topics evolve over time (hence the "dynamic" adjective). Conveying the results of topic modelling can be challenging since the topics often do not lend themselves naturally to meaningful labelling. The volume of real world (global) social media also means that millions of topics can be ongoing at any given time and the relationships between them can involve hundreds of dimensions and relationships that continually emerge. The popularity of topics is itself subject to change over time and reflect the pulse of what is happening in society at large. In this paper, we propose a spatialization technique based on open-source software that reduces the intrinsic complexity of dynamic topic modelling results to familiar topographic objects, namely: ridges, valleys, and peaks. This offers new possibilities for understanding complex relationships that change over time whilst overcoming issues with traditional topic modelling visualisation approaches such as network graphs.

1. INTRODUCTION

The Australian Research Data Collection (ARDC) funded Australian Data Observatory project (ADO - www.ado.eresearch.unimelb.edu.au) has been collecting tweets and other social media posts (Instagram, Reddit, YouTube, Flickr, etc) related to Australia since June 2021. Through the use of the new Twitter Academic Research access, the project is now harvesting approximately ten million social media posts per month. The social media posts are stored and analyzed daily using the deep learning BERTopic package (Grootendorst, 2022). A BERTopic output is stored and subsequently made accessible through a Representational State Transfer (ReST)-based application programming interface (API) supporting different clients, e.g. through Jupyter notebooks and an associated web application. The intended audience of the platform is broad and includes social scientists, humanities researchers, linguists, data journalists and big data researchers. The goal is to support data exploration at scale and overcome the smaller scale cottage industry of social media research that has hitherto been the norm across academia in Australia. The scale of data however means that standard approaches for data analytics and especially data visualisation need to be revisited. There is simply so much data that the "big picture" of what is happening is often lost in the noise. The goal of this paper is to present a novel approach for visualisation of evolving social media topics based on the metaphor of topography.

Topic modelling is often used to understand what is happening online in social media platforms. A topic model is a form of statistical model used to discover abstract topics of discussions that occur in a collection of documents (social media posts). It can be used to discover semantic structures in text corpus. Intuitively, given a set of documents, one would expect particular words to appear in the documents more or less frequently. For example, "player" and "team" may appear more often in documents about sports - even if a given (single) document does not include these two terms or does not specifically include the term sport. These associations give rise to patterns in the text that can be used to identify given topics (sports, team, player, ...) etc.

There are different models that have been applied for topic modelling. Latent Dirichlet Allocation (Blei et al., 2003) applies a statistical approach, whilst others rely on natural language approaches such as BERT (Devlin et al., 2018). Based on experiments, we found the latter gives results that are more meaningful for large-scale collections of continually growing and evolving social media data. Models that rely on language models represent topics as vectors in a high-dimensional space (at least 384 dimensions for BERTopic (Grootendorst, 2022)), which presents a challenge when it comes to visualizing the results of any associated analysis. The representation of topics themselves can take different forms, but they are often presented using 2D visualizations, such as circles with size proportional to topic popularity and with position related to the similarity between topics (Karpovich et al., n.d.), or other representations such as Sankey diagrams (Murakami et al., 2021). For the ADO project (ADO, 2022b) we implemented 2D visualization to show the results of dynamic topic modelling: a graph composed of topics as nodes (circles of size proportional to the number of documents) and edges connecting topics, belonging to different days, that share a given number of most significant terms (see Figure 1).

However, such representations do not convey the evolution of topics over time satisfactorily. For example, using a 2D space to represent topics relies on the use of dimensionality reduction techniques, such as UMAP (McInnes et al., 2018) to reduce the number of dimensions of a vector to two, but this does not

* Corresponding author

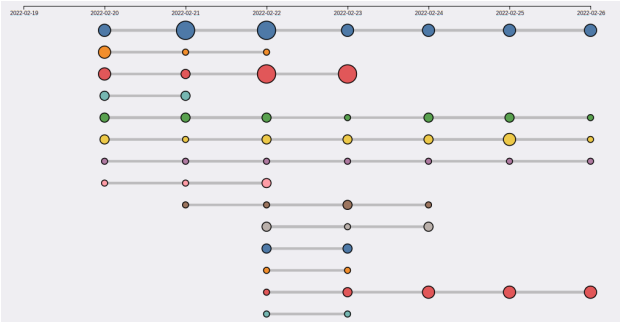

Figure 1. Example of dynamic topics visualization using a 2D graph

take into account the evolution of topics over time to cater for the dynamic aspect of topic modelling and emergence of new topics and topic relationships.

Spatialization (Fabrikant, 2017) is a technique that uses spatial metaphors to aid cognitive tasks. It has been a research field since the early 1990s. It can be used to construe vast amounts of information by reducing them to a physical landscape that can be visibly traversed with relative ease.

In this work, we consider spatialization of topics in a 3D space where:

- the X-axis represents time based currently on one-day intervals;
- the Y-axis is the distance (one minus cosine similarity between the vectors representing two topics) between topics posted on the same day, and
- the Z-axis is a measure of the topic popularity.

With this approach, a topic is therefore reduced to a single point in a 3D space, and the interpolated surface constructed out of these points becomes a landscape with peaks, ridges, and valleys. More precisely, the “valleys” represent less popular topics, while “peaks” represent the more popular ones and flat surfaces indicating the topics with average popularity. The dynamic (temporal) aspect is represented by the development of ”ridges” and ”valleys” along the X axis. Bringing in the third dimension allows to represent the relationships between topics and the evolution of their popularity at the same time.

This 3D landscape naturally aids the end-user in understanding complex highly dimensional data at a scale and volume that would otherwise be impossible. The formation of mountain ranges or valleys related to mainstream topics such as COVID-19 through to geopolitical events such as the invasion of Ukraine provides a finger on the pulse of what is being discussed at scale by the broader population across the social media landscape. To allow users to understand what a topic is about, the point of each topic on the surface has a label that shows its most representative terms - currently this is set to show the five most popular terms.

2. DATA AND METHODS

The construction of the topographic surface is based on the dynamic topic modelling procedure provided by BERTopic. This is based on the transformation of documents, i.e., social media posts, into vectors (called *embeddings*) of a high-dimensional space, and on the subsequent grouping of related documents into topics using a hierarchical clustering technique.

Since the topic modelling is based on the BERT language model, it is able to place terms into their semantic contexts, thereby differentiating between synonyms and yielding better classification results overall.

Once the topics are computed, X, Y, and Z coordinates are assigned to each topic, making it into a point in a 3D space. The last step is then the computation of a continuous surface from said points to an associated topographic representation. In more detail, the topic modelling and topographic surface building procedure is composed of many different steps, summed up in the following list:

- social media post harvesting to establish a corpus (set of documents);
- corpus parsing and cleaning;
- systematic sampling of social media posts to reduce the overall corpus size;
- merging of individual social media posts into documents composed of related posts also referred to as (*conversations* or *threads*);
- document embedding through BERTopic;
- dimensionality reduction of the embedding vectors to improve clustering, as high dimensionalty tends to shorten relative distances between topics, making clustering less discriminating;
- clustering of documents into topics based on the distance between the dimensionally-reduced embedding vectors;
- hierarchical clustering of topics into more general topics to reach the given minimal size (Joachims, 1997);
- assignment of labels to the topics, based on a class-based TF-IDF measure of terms frequency (Joachims, 1997));
- subdivision of topics based on the date of the tweets, allowing for tracking of document frequency per-topic or per-day etc;
- dropping of the topics composed of all outlier documents, i.e. documents that cannot be grouped into topics;
- further dimensionality reduction of topic embeddings since each topic is represented as a high-dimensional vector and can be reduced to one dimension to spread the topics across the Y axis such that topics with similar content are closer on this axis);
- computation of the X coordinate as time (days);
- computation of the Z coordinates as number of documents of each topic per-day;
- standardization of X and Y axes in the [0, 1] interval;
- computation of the surface on a square grid where the Z coordinate is computed using a Kernel Density Estimator with a ”skewed” distance, to compensate for the larger distance between points on the X axis as compared to the one along the Y axis;
- re-scaling all three axes to improve the aesthetics of the resulting surface, and finally
- visualization of the grid as a topographic surface.

The data set used in this analysis in this paper was conducted using Twitter data harvested and stored using the ADO system over a period of 20 days between 20th of February to 11th of March 2022. This comprised about 3.2M conversations (discussion threads) posted by Twitter users that were either located in Australia or that mentioned Australia in their Twitter profile. The harvesting excluded retweets to avoid document

repetitions and to minimize, to the extent possible, content that would likely be produced by "bots".

The tweets where extracted from the ADO CouchDB database and went through a syntactic cleaning process, which consisted of:

- dropping tweets that were not tagged as being written in English (this does not exclude that some tweets, written totally or partially in languages other than English, could have been retained);
- tweets were grouped in *conversations* (threads), to increase the size of documents in the corpus and improve the quality of the semantic document embedding of BERT;
- URLs, numbers, usernames, and hashtags were dropped;
- all terms that were not recognised as nouns (adjectives, verbs, etc.) were dropped;
- short tweets (fewer than ten terms) were dropped, and
- terms were "stemmed" using the Porter Stemming Algorithm (Porter, 2006) i.e., the inflected or derived words were reduced to their word stem -"plays" and "played" were reduced/stemmed to "play" and so on.

The cleaned corpus was then fed into BERTopic for dynamic topic modelling, yielding a a set of vectors representing topics across the whole time interval considered, i.e., the entire corpus was used to extract topics. The topics were the same from the first to the last day, but with different popularity (measured as number of tweets related to that topic per day). To avoid having too many topics, a minimum number of posts (1,000 across all days) was imposed for a topic to be retained, hence the least popular topics were merged.

Given the computational resource (memory) constraints where only a 128GB RAM virtual machine was available, the 3.2M corpus had to be systematically sampled. This was achieved by halving its size (every odd tweet in order of reading from the corpus file was discarded).

It is noted that BERTopic uses UMAP internally before hierarchically clustering topics together, as the high number of dimensions generated by document embeddings does not lend itself well to clustering, i.e., distances between clusters tend to converge to the same distance in high-dimensional spaces. Therefore, even with a relatively high number of documents, the topics generated (and their count) was subject to fluctuations from run to run. While this artefact can be eliminated by just initializing the UMAP random state to the same number, we identified that it would be better to expose the stochastic nature of the algorithm and assessing its relevance instead. As an example of this, two different runs of BERTopic on the same resampled 1.6M-document corpus gave topics that had the same or similar most important terms (Table 1 and Table 2). Here the size of topics is expressed as the number of tweets with different, but similar values. It must be noted that some topics include "miscellaneous" ones whereby the most popular topic in either Table 1 and Table 2 may well be expressed in languages other than English (the eighth topic of Table 1), which can cause some noise in the results.

Once the topics were obtained, their vectors were clustered using HDBSCAN (McInnes et al., 2017) and UMAP to obtain a set of topics ordered along the Y axis, with the distance between any pair *roughly* proportional to the distance of the full-dimensional original vectors. Roughly is italicized in the

Size	Terms
71607	and
12994	covid, vaccin, death , infect, case
23384	flood, rain, wate, weather, storm
10782	food, dinner, meat, chicken, pizza
12966	goal, player, season, team, club
26316	hai, ang, aku, bir, ako
23533	putin, russia, russian, ukrain, nato
14084	que, por, lo, para, con
21175	song, music, album, listen, sound
12868	women, gender, intern, woman, men

Table 1. Most popular topics of the first sample run - sorted by terms.

Size	Terms
13125	covid,vaccin,death,case,infect
23180	flood,rain,water,weather,storm
28597	food,coffe,dinner,chicken,beer
27962	hai,ang,aku,bir,ako
21333	putin,russia,russian,ukrain,nato
14281	que,por,lo,para,con
14562	season,player,goal,match,team
21082	song,music,album,listen,sound
12536	women,gender,intern,woman,men
71608	yer,mabunt,of,herd

Table 2. Most popular topics of the second sample run - sorted by terms.

previous sentence due to the stochastic nature of UMAP, in common with other dimensionality reduction algorithms. There is a substantial reduction of scale (from 384 dimensions to 1). Therefore, as a consequence of this stochastic nature, different runs of the same UMAP process can give different results, possibly yielding more accurate results for a subset of topic pairs, and less accurate results for other pairs. As noted, the Y axis values were then scaled to fall in the [0, 1] interval.

It was identified that one parameter that had a considerable impact on the UMAP dimensionality reduction was the *number of neighbours*. This guides the algorithm towards a more global view of the data structure (larger values), or a more localized one (smaller values). We experimented with different values over many iterations, e.g., number of neighbours equal to 2 (BERTopic default for topic visualization), 5, 10, and 30, over 100 iterations each. It was identified that the resulting average values were more concentrated with small parameter values. In the interest of having better topic separation. we opted for a value of 10 for the number of neighbours parameter. As per the distance measure between topics, we opted for the *cosine similarity* (Manning et al., 2008, p.120-123) or more specifically *one minus cosine similarity*), in keeping with the metric BERTopic uses internally.

Utilising the Y axis coordinates of points, the procedure assigns values on the X axis (time) according to the day of posting of a given tweet in the [0, 1] interval (i.e., if there are 10 days in the interval, all the tweets of the 5th day get an X value of 0.5).

The Z value (topic popularity) is proportional to the number of tweets on a given topic for a given day.

The result of this procedure is a collection of 3D points in the domain:

$$x \in [0,1], y \in [0,1], z \in [1, \max\{n\}]] \quad (1)$$

where *n* is the number of tweets per topic, per day.

To balance the X and Y coordinates with Z, and to get a representation that is more landscape-like, a re-scaling can be applied to all axis: we found that the best results were obtained by re-scaling X and Y by 1,000 and leaving Z unchanged; however, different data and time intervals may need different scaling factors.

A sample of the results of this process are shown in Figure 2, where:

- topics are repeated across days;
- topics that are more closely related appear closer on the Y axis;
- the distance between points on the X-axis is noticeably larger than the distance on the Y-axis since there are more topics than time intervals.

Figure 2. Five topics on two consecutive days displayed as 2D points with the five most significant terms of each topic shown as point labels.

The difference in the average distance between points on the X- and Y- axis is due to the number of different topics on the same day and the number of days, e.g., 100 topics over 10 days. While there are ways to correct this, current approaches are limited, e.g., having the same number of days as topics runs into processing time constraints (for instance, a 10-day -full, not re-sampled- dataset required a computation that ran for 20 hours on a 128GB RAM, 32 vCPU virtual machine), whilst using a distance between days (X axis) equal to the average distance between topic on the Y axis yields points that are disposed along a very tall and narrow strip, making it unsuitable for visualisation, and knowledge extraction and understanding of the online discourse.

One solution to this is to compute a surface that bridges the gaps between topic points along similarity (Y axis) and time (X axis). Different interpolation methods have been explored in this work, and the quartic Kernel Density Estimator (KDE) was found to give the more aesthetically pleasing results. However, the use of KDE has its own challenge based on the different average distances between points along the two axis. The solution to this was to use a skewed Euclidean distance that "shortens" the distance on the horizontal axis (time) by a factor proportional to the ratio between the number of topics *n* and the number of time intervals *t*, i.e., the number of topics per time interval. Therefore, the Euclidean distance between points p and q becomes:

$$d\left(p,q\right)=\sqrt{\left(q_x-p_x\right)^2/\left(n/t\right)+\left(q_y-p_y\right)^2} \qquad (2)$$

A specific quartic KDE is defined by the maximum distance for the value associated with a point (the Z value) that has an influence on the surface. Different values yield visually different results. For instance, when the maximum distance is less than one time interval, the surface starts to appear "corrugated" (Figure 3), while a larger value gives more uniform results (Figure 4), which may hide local "dips" or "peaks" in topic popularity. In addition, larger maximum distance values bring higher Z values, which can be be countered by re-scaling the output surface.

Figure 3. A KDE-interpolated surface using a small maximum distance.

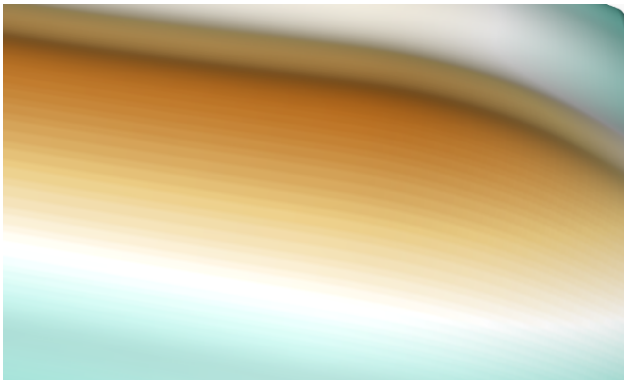

Figure 4. A KDE-interpolated surface using a large maximum distance.

The points themselves are written in a GeoJSON for visualization, and the surface is written as an ESRI ASCII Grid. The visualizations in this paper were all done using QGIS (QGIS, 2022) and its QGIS2ThreeJs plugin (Akagi, 2022) for 3D rendering. These include the visualization of the top-5 terms of each topic as labels draped on the surface. The color scale used is single-band pseudo-color, with five linear interpolated classes (from lower to higher values: dark green, green, white, light brown, dark brown)

3. EXAMPLE RESULTS AND DISCUSSION

We tested our methodology on the 20-day Twitter data set described above. The resulting surface exploration gave some interesting insights. For instance, the popularity of topics related to the "The Batman" movie including movies in general, books, and streaming services peaked on the 4th of March 2022 with the release of "The Batman" movie. This included considerable chatter a few days prior to the movie release and a sudden drop afterwards (see Figure 5).

Another insight was the "ridge" related to everything to do with "politics", which was the most popular topic during the period

and was strongly correlated with topics about the job market (Figure 6).

A more general view of the visualization (Figure 7) shows how closely it resembles a physical landscape, although one with rather steep ridges and a corrugated-looking crests of the various features. It should be noted that these also reflect artefacts of the specific set of interpolation parameters used as much as the results of topic modelling and the associated dimensionality reduction.

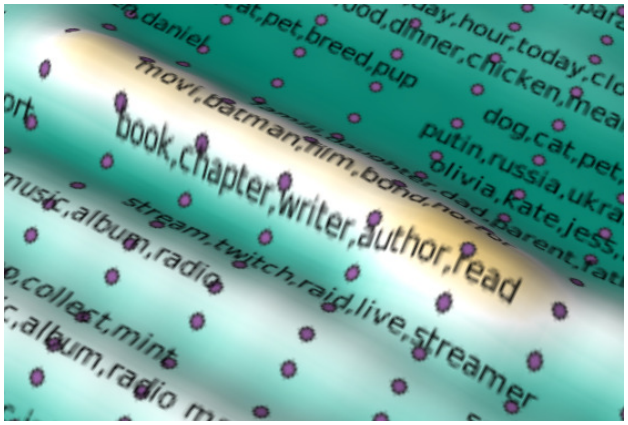

Figure 5. Batman Peak visualization.

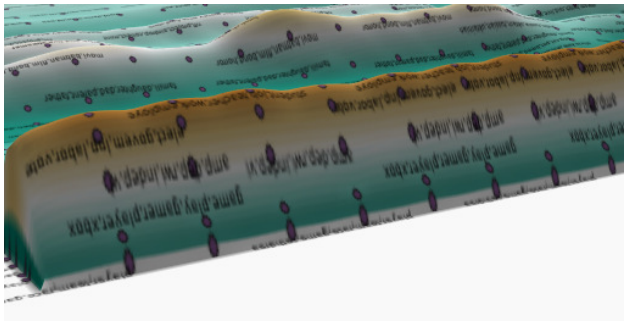

Figure 6. Politics Ridge visualization.

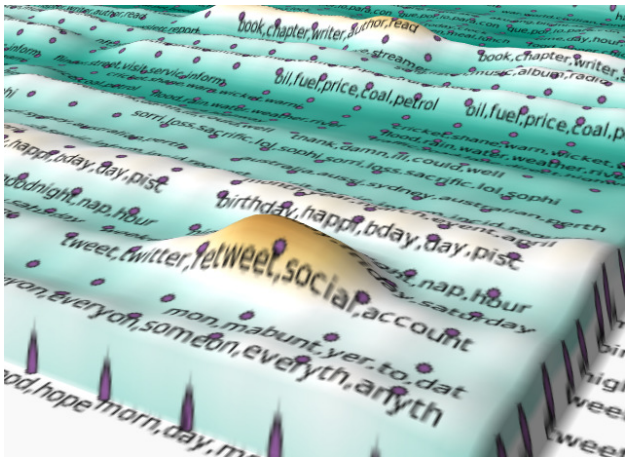

Figure 7. Extract of 20 Days Landscape visualization

4. CONCLUSIONS

The proposed topographic visualization allowed us to gain insights by examining the way topics are correlated to each other and the relative popularity of each topic over time.

Whilst we found that the proposed visualization conveyed information that was both expressive and intuitive (Figure 8), it relied on two inter-correlated parameters: the maximum distance used for KDE and the Z-axis re-scaling. Increasing the value of the maximum distance smooths the surface but increases the Z values, subsequently giving rise to Z values that are hard to explain and have visually unpleasant results. We settled on both values being equal to one (i.e. maximum distance being equal to a given (fixed) unit time interval and no re-scaling of the topic popularity). However these are somewhat subjective values and may well be different when other data sets are considered. More work is needed in this area to compute an optimal set of parameters exclusively from the available data, leaving the subjectivity out of visualization.

In addition, interpolation techniques other than KDE may be tried, although we already attempted and discarded approaches such as Kriging (Cressie, 2015), an interpolation originally proposed for geo-statistics to model the surface of aquifers or ore seams, but subsequently used to model other phenomena. We considered Kriging for improving the aesthetics of the visualization, as this technique yields smoothed surfaces that we perceived to improve the user experience. However, Kriging relies on spatial auto-correlation and on the notion that points are independently sampled, while our points represents topics that are not independent: intuitively, a topic in a given location ”attracts” document in the vicinity, resulting in fewer documents for topics nearby, hence negative spatial auto-correlation.

Perhaps the most valuable result from this visualization technique is the critical appreciation of topic modelling results it allows. Since topic modelling and dimensionality reduction are stochastic algorithms, they tend to give different results in different runs over the same data set. This fuzzy aspect of topic modelling is difficult to grasp just by looking at the topic modelling results using a 2D visualization, but it becomes obvious when looking at it as a ”landscape”. Thus sometimes the ridges are not as high, or they appear in different places in different runs, and this gives an immediate appreciation regarding the robustness of the results.

A serious obstacle to our exploration of topic modelling is the sheer time it takes to run such computations on a sizeable data set such as the one provided by the ADO project, especially when runs are to be repeated to yield more stable results.

Therefore, our next objective is to:

- split the procedure in different steps and save the intermediate results in the interest of modularity;
- have runs of the dimensionality reduction step (the one used to compute the Y coordinate values) execute in parallel on a cluster of computers to decrease computing times;
- aggregate results of different runs into a single visualization.

The overall goal is in reducing the inherent fuzziness of topic modelling and dimensionality reduction to obtain more stable results through iteration and subsequent aggregation.

The software used to compute the topic models and the visualization has been published at (ADO, 2022a) and is available under an Apache Software License 2.0.

The original corpus (text of the tweets composing the data set) used for this paper cannot be made publicly available due to the privacy safeguards the ADO project adopted, and further restrictions connected to our application for the Twitter API Academic Research access. However, the authors can, on demand, make the data set tweet IDs available to third parties that can then proceed to download themselves the tweets text using the Twitter API and replicate our results. Information on how this is achievable is given on the ADO website (ADO, 2022b).

REFERENCES

ADO, 2022a. Ado-topoviz. https://github.com/lmorandini/ado-topoviz.

ADO, 2022b. Ado website. https://www.ado.eresearch.unimelb.edu.au/.

Akagi, M., 2022. 3d visualization powered by webgl technology and three.js javascript library.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, 993–1022. http://portal.acm.org/citation.cfm?id=944937.

Cressie, N., 2015. *Statistics for Spatial Data*.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Fabrikant, S., 2017. Spatialization. https://doi.org/10.1002/9781118786352.wbieg0812.

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Joachims, T., 1997. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 143–151.

Karpovich, S., Smirnov, A., Teslya, N., A., G., n.d. *20th Conference of Open Innovations Association (FRUCT)*.

Manning, C. D., Raghavan, P., Schütze, H., 2008. *Introduction to Information Retrieval*. Cambridge University Press.

McInnes, L., Healy, J., Astels, S., 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11), 205. https://doi.org/10.21105/joss.00205.

McInnes, L., Healy, J., Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction.

Murakami, R., Chakraborty, B., Y., S., 2021. Dynamic topic tracking and visualization using covid-19 related tweets in multiple languages. *2021 International Conference on Artificial Intelligence and Big Data Analytics*, 16–21.

Porter, M., 2006. The porter stemming algorithm. https://tartarus.org/martin/PorterStemmer/.

QGIS, 2022. Qgis geographic information system.