Glance: Visualising Reddit with Topics Over Time

Team 21: Mehmet Suat Gunerli, Asha Gutlapalli, Dipendra Singh Mal, Yashwant Singh, Anshit Verma, Jaylen Alexandra Williams

1 Introduction

As more and more information is collected on the internet, it gets harder to keep up with all the news, developments and discussions pouring in across a multitude of domains including technology, politics, health, sports, etc. and follow a topic of interest. It is even harder to translate that information into valuable insights. With Glance, our motive is to detect and keep track of emerging topics on Reddit, while extracting more useful and hidden information by leveraging user feedback and other linked content so that individuals or organizations are able to stay on top of trends, plan ahead and make data-informed decisions.

2 Problem Definition

People nowadays talk about various subjects on social media platforms daily to share their thoughts, express their feelings, find answers to their questions and debate with one another. The objective of this project was to visualize different topics of discussion within Reddit posts and comments, and depict how those topics evolve over time. Our goal was to build an interactive visualization tool for users to find relevant topics that pique their interest, while simultaneously keeping an eye on other emerging topics and trends as they evolve over time.

3 Literature Survey

3.1 Topic Models

Latent Dirichlet Allocation (LDA) -one of the popular topic modeling algorithms- was used in conjunction with k-nearest neighbors (KNN) algorithm to group information as well as identify whether that information was genuine or not. However, this approach failed to perform well on conventional datasets despite categorizing the data into distinct categories [13]. Identifying the right hyperparameters in the model and optimizing them also proved difficult [4]. Another widely used topic model is the Correlated Topic Model (CTM). CTM was also used to assess the relationships between discourse materials. It tends to perform well when the data it was previously trained on is similar to the data that it is tested on with respect to the domain. Yet, it is extremely sensitive to user-specified parameters like the number of topics that the article can be assigned to [16]. Structural topic modeling is an R-based topic modeling approach used for clustering text documents and these clusters were illustrated with the help of word clouds. Word clouds express the significance of words in a particular topic. Nonetheless, a better visualization method could have been used to illustrate the relationship between different topics [15]. To perform semantics-based dynamic interest finding, a topic modeling approach is implemented to implicitly capture word–word, word–author and author–author, word–time, and author–time relationships, simultaneously [6].

The time decay function coupled with LDA will indicate topic strength along with topic similarity. However, discrete time intervals in a time decay function may fail to capture time dependence of word occurrences [21]. Controlling topics in a way that makes it constant will deteriorate the performance of the topic model and produce unnecessary topics [19]. On-Line LDA has a beneficial effect on small subsets of data as it is efficient in both memory and speed [2]. In addition to the topics themselves, the actors of the topics are essential to take into account when building a topic model as not doing so will lead to a lot of contextual information being lost [8]. Utilizing the update algorithm along with LDA renders it possible to update the models in real time without needing to recalculate the whole model, thereby improving efficiency. Though weighting and random sampling of previous documents as possible tools would serve as advancements in topic modeling performance, it fails to propose a definitive process or a model [17]. Overall, LDA delivered relatively more meaningful extracted topics and obtained better results than other topic models [1]. In general, topic modeling is functional for comprehensive quantification and practical representation textual data [14][10]. Nevertheless, topic modeling must evolve to adapt for different document lengths [5].

3.2 Clustering Algorithms

K-Means clustering is one of the most popular unsupervised machine learning algorithms in literature and industry. This model was used to group both visual and textual documents that are alike, thereby bringing in additional dimensionality to the model to help improve its decisions. However, a limitation of the K-Means algorithm is that the number of clusters must be specified prior to carrying out the clustering. When it comes to topic modeling, the number of clusters cannot be determined before the model is run. [12].

3.3 Other Methods

Newman's modularity algorithm is a hierarchical algorithm implemented to measure the degree of interconnection within a group and intraconnection between different groups. It is instrumental for dimensionality reduction. However, the algorithm is not deterministic which can lead to the formation of different groups per each run [20]. Excess Correlation Analysis (ECA) is another efficient method to recover the various parameters in a topic model like LDA. This improves LDA in terms of time complexity and arrives at the resultant topics faster. Yet, CSA is unstable as it needs randomization to separate the singular values when computing the Singular Value Decomposition (SVD) [3]. There is a unique approach called closed frequent keyword set to form topics. This approach provides a natural method to cluster the documents into hierarchical and overlapping clusters using the topic as a similarity measure [18].

4 Proposed Method

4.1 Intuition

Our proposed method uses text classification to extract high frequency topics on Reddit. We successfully incorporated a time component into topic modeling, displayed using interactive graphs and cluster visualizations to allow users to identify and follow popular or trending topics over a given time window. Our approach is successful at allowing users to see what kinds of topics stay relevant the longest and if certain times of the year have similar topic popularity trends.

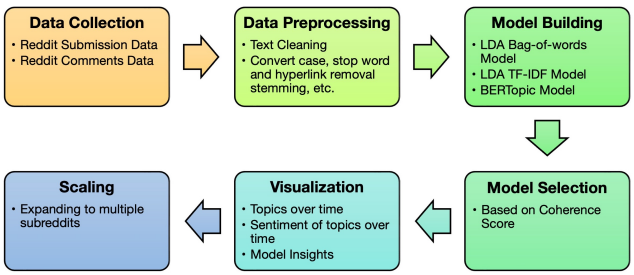

Figure 1: Schematic Diagram of the Proposed Method

4.2 Detailed description

Latent Dirichlet Allocation (LDA): LDA, a popular topic modeling technique, was explored with different feature engineering methods to identify significant topics from the collected Reddit data. Topic modeling, on the other hand, is an unsupervised statistical model capable of mining topics using frequently occurring words in the given documents. These documents can be filtered based on the minimum number of words, the content of the documents, the popularity of the documents, etc. Before feeding these documents to the LDA model, it is crucial to perform data pre-processing. Each document is assigned a topic based on its distance from the documents of all topics. If the distance is above a certain threshold, then the document is classified as a specific topic. Then, LDA generates a matrix comprising probabilities of words against topics. For a word, the topic with the highest probability indicates that the word is most likely to belong to that particular topic. After filtering suitable documents, lowercase conversion, removal of stop words and punctuation, tokenization, and lemmatization were done to create a cleaned word list for the subsequent feature engineering methods.

Bag-of-words: Bag-of-words is a feature engineering technique for text classification. This model turns a given text into vectors that counts the frequency of word appearances. While Bag-of-words is simple to implement and compute, a downside of this technique is that it does not take into consideration the ordering of words in reference to other words in the document. This can lead to most frequent topic results of words not being correlated to one another. A proposed approach to combat this is using n-grams to vectorize text as a predefined range of text length [11].

TF-IDF: TF-IDF (Term-Frequency Inverse Document Frequency) is a feature engineering technique that is a composite function of term frequency and inverse document frequency. Term frequency is the number of occurrences of a term in a document over the total number of terms in the document. Inverse document frequency is the log of the total number of documents over the number of documents with a distinct term. Term frequency captures the frequency of a word occurring in a particular document and inverse document frequency captures the distribution of the word over all documents. Using inverse document frequency in conjunction with term frequency ensures that the word is not only frequent but possibly unique to a particular document. For example, though the word "the" is common in a document, it is also common in all other documents. Thus, TF-IDF does not consider this word important in that document. TF-IDF gives out a matrix comprising weights of words against documents.

BERTopic: BERTopic was explored as it leverages the power of transformers and TF-IDF to generate topic representations. In this approach, a document is first represented by a word embedding which was learned by a pre-trained language model, then dimensionality re-

duction using UMAP is performed to optimize clustering and, at last, HDBSCAN is used for clustering. A class-based TF-IDF alogrithm is used to generate importance of word to a particular topic instead of a document. To account for the temporal nature of the documents, the global representation of topics are learned first. Then, the local representations are learned in which the topic c-TF-IDF vectors at time t depend on those at time t-1 [7]. BERTopic was used on the pre-processed titles obtained from the Reddit posts as described in the Data Mining section of the report.

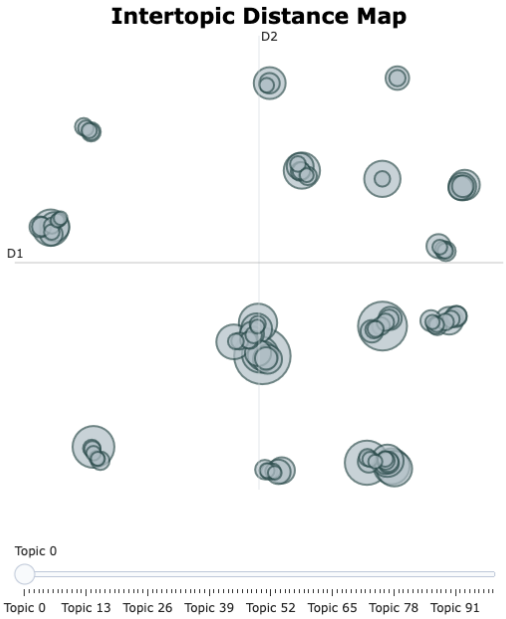

Figure 2: Intertopic Distance Maps

Figure 2 shows the intertopic distance where the centers are determined by the Jensen–Shannon divergence and then using multidimensional scaling to project it onto a 2D plane. It shows how topics differ from each other and the size of the circle indicates the prevalence of a given topic.

Figure 3: Topic representations of the top 4 topics

Figure 3 shows the global distribution of words for the top 4 topics. We can see that the model is able to capture the most important words so that the topic makes sense. As shown in Figure 3, Topic 0 is about the Coronavirus and its spread, Topic 1 is about the Israel-Palestine conflict and what is happening on the West Bank, Topic 2 is about the Russian invasion of Ukraine and Topic 3 is about the Coronavirus vaccines.

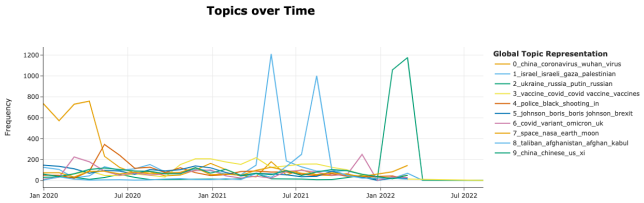

Figure 4: Topics over Time

Figure 4 shows the frequencies of a number of topics over a period of time which can help the user determine the prevalence/popularity of topic at any given time and track how it changes.

Words	Timestamp
coronavirus, uk, virus in, confirmed in, cases	2019-12-31 00:19:13.393999872
uk, coronavirus, covid, cases, confirmed	2020-02-02 21:23:48.200000000
covid, coronavirus, uk, cases, variant	2020-03-06 18:47:28.400000000
covid, sweden, coronavirus, deaths, uk	2020-04-08 16:11:08.600000000
sweden, covid, coronavirus, deaths, uk	2020-05-11 13:34:48.800000000

Figure 5: Evolution of topics over time

Figure 5 shows how a given topic evolves over time. As seen in the figure, the topic first emerged with Coronavirus cases in the UK, then the word *variant* gains traction in March 2020, followed by the word *deaths* in the next month. The evolution of this particular topic summarizes how the Coronavirus situation changed in the UK over a period of time.

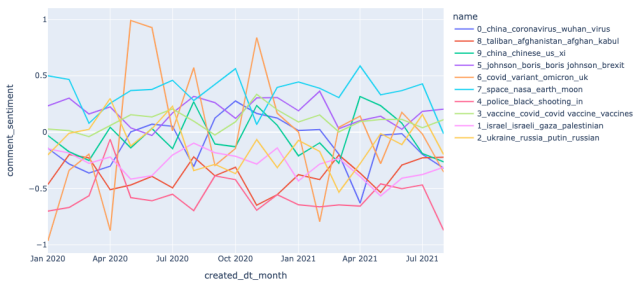

Figure 6: Topic sentiments from comments

Figure 6 shows sentiment scores for each topic based on user comments. In the lack of labeled data for sentiments, we used *vaderSentiment*, a lexicon and rule-based sentiment analysis tool that is specifically tuned to analyze sentiments in social media. [9]. Here the values 1, 0, and -1 correspond to positive, neutral and negative sentiment, respectively. We can see how a topic related to space exploration and NASA trends positively over time, whereas a topic related to Taliban and Afghanistan remains negative throughout.

4.3 User Interface

Our user interface was designed to identify trending topics on Reddit easily and quickly. The initial display shows the top 20 topics for each of the three consecutive months. The user interface allows for filtering on specific subreddits (r/worldnews, r/movies, r/politics) from a drop down select menu and for identifying topic comparisons and trends among other months of the timeline (January 2020 to September 2022). The size and color shade of the topic modeling bubble represents the frequency of the topic in a certain month, where larger, darker bubbles (topics) have a higher frequency.

By hovering over a topic, a user can also see what words are associated together for the topic, and the sentiment level for that specific topic.

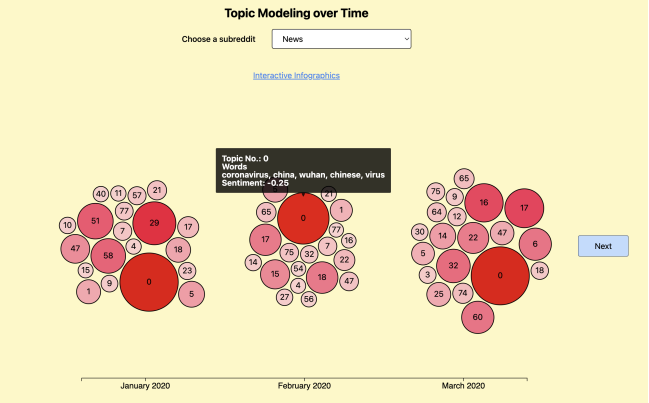

Figure 7: Topic Modeling User Interface

By clicking on a topic bubble, a user can identify the inter-connectivity of topics and how the trend of a topic is changing over time. One can also see the changing sentiment levels on this screen, where topics closer to -1 (bottom of screen) have lower sentiments and topics closer to 1 (top of screen) have higher sentiments (shown in Figure 8).

An example of topics changing each month in a year is under the "News" subreddit. Topic 17's sentiment has a negative sentiment score of -0.12, -0.16, and -0.10 in January, February, and March 2020 respectively. The negativity of this topic becomes even more negative in February but then becomes pretty much neutral in March than in January. The top five words were "coronavirus", "confirmed", "case", "cases", and "of coronavirus" in January 2020. The top five words were "coronavirus", "cases", "death", "death toll", and "toll" in February 2020. The top five words were "coronavirus", "cases", "coronavirus cases", "toll", and "death toll" in March 2020. The top words change to "death toll" and "toll" in February and March compared to "cases" and "confirmed" in January. This shows that corona deaths were concerning since February 2020 compared to confirmed cases in January. The number of Reddit posts regarding this topic also increased in March 2020 compared to January 2020.

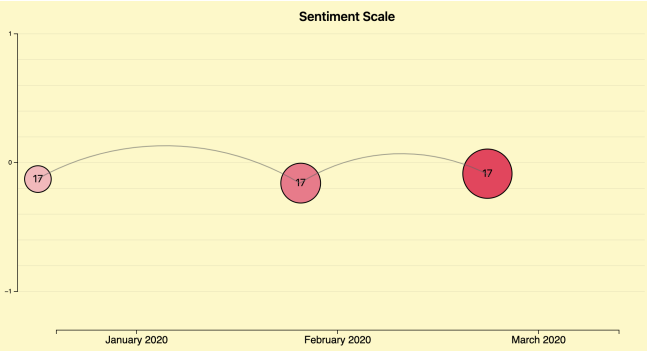

Figure 8: User Interface Topic Trend and Sentiment Page

At the back-end of the interface we choose BERTopic as the topic modeling algorithm to perform on the Reddit data because of its higher coherence score shown in section 6.3 of Observations and Evaluation.

Towards the top of the display, under the filter option, there is a link to another page with the interactive BERTopic visualizations for a user to understand the topic modeling algorithm. These visualizations are mentioned in section 4.3 Detailed description and are the interactive versions of Figures 2, 3, 4, and 6.

5 List of Innovations

1. Using Reddit posts for topic modeling and showing how the topic is trending and evolving over time.
2. Embedding linked content sourced from associated URLs and user comments within the corpus.
3. Using sentiment analysis to analyze user response and capturing the relationship between user response and the relevancy of trends.
4. Interactive visualization tool allowing users to easily and quickly see trending topics rather than clicking through large amounts of Reddit posts.

6 Experiments and Evaluation

6.1 Test Bed Description

1. What was the most trending topic for a given year on Reddit?
2. How long do popular topics remain popular?
3. Are there periods of the year where topics have a higher frequency?
4. Does including comments in addition to titles and posts change the results of the model?

6.2 Design of Experiments

1. Using the LDA/BERT Model:
   (a) Testing scalability and generalizability of the proposed topic model and identifying whether it is possible to source content from multiple subreddits to build the corpus.
   (b) Identifying the best performing topic model among the proposed alternatives using coherence score measures.
2. Using Visualizations:
   (a) Detecting similarity and parallelisms among topics and trends sourced from bodies of posts, URLs, and user comments using inter-topic distance maps and "Topics over Time" charts.
   (b) Identifying the top n most popular topics identified by the topic model using frequencies within a specific time window and visualizing how those topics trended over time.

6.3 Observations and Evaluation

6.3.1 Topic Modeling Algorithm Evaluation

For the three proposed methods, we use coherence score to figure out the optimal model. Coherence measures score a single topic by measuring the degree of semantic similarity between high-scoring words in the topic.

Looking at Figure 9, we conclude that on the basis of coherence scores, the BERTopic technique does the best job for topic modeling on our data.

Model	Best Coherence Score	Num Topics
Bag-of-Words LDA	0.3	8
TF-IDF LDA	0.37	8
BERTopic	0.49	10

Figure 9: Model Comparison

6.3.2 BERTopic Subreddit Evaluation

Our next step was to use BERTopic to model for the three subreddits we had chosen and the results are shown in Figure 10.

Subreddit	Coherence Score	Num Topics
r/movies	0.44	83
r/worldnews	0.41	46
r/politics	0.26	391

Figure 10: Final Models

6.3.3 Usability

Our interface has high usability because it doesn't require the user to input any information and the visualization is easy to use and understand, allowing for few to no errors in obtaining information. Many of the current topic modeling interfaces and applications have cluttered information and don't provide easily accessible information on what algorithm was used to construct the application. By creating a spacial interactive visual with the option to access the BERTopic information by clicking a link, a user can identify a simple model trend or access more involved topic trend data.

6.4 Interactive Infographics

Moreover, each subreddit has an interactive infographics dashboard. This dashboard displays the sentiment evolution of comments about topics, intertopic distance, the frequency of topics over time, and topic word scores. For example, in the "movies" subreddit, comments about Christopher Nolan were the most positive in January 2022 since 2020. Understandably, the topic of Christopher Nolan and Batman form the same cluster and are distanced from other clusters that talked about Indiana Jones or Spider-Man. People talked about Christopher Nolan towards the end of 2020 and the start of 2021 compared to other months from 2020 to 2022. People talked about his movies like Inception, memento, and time the most in the topic of Christopher Nolan. In this way, each topic can be studied in detail in terms of its frequency and sentiment over time, the variability in topics, and the most significant sub-topics in each topic using line graphs, scatter plots and bar charts.

6.5 Data Mining

We used the Pushshift API to source the data from Reddit across three subreddits in r/worldnews, r/movies, r/politics. Roughly 1.5 GB in size, our dataset comprises over 300,000 Reddit posts, 2,250,000 comments and related content

sourced from linked URLs using the newspaper3k library, sent between January 2020 and September 2022.

	id	subreddit	created_utc	url	title	num_comments	comments	url_content
0	eib002	worldnews	2020-01-01 00:00:08	https://www.reuters.com/article/us-india-citiz...	Thousands of Indians ushered in the New Year b...	17	[deleted]...Because most Indians are not Musli...	NEW DELHI (Reuters) - Thousands of Indians ush...
1	eib9ut	worldnews	2020-01-01 00:21:23	https://bgr.com/2019/12/31/china-pneumonia-sic...	A scary unidentified virus is spreading in China	70	[removed]...Ah, a fellow time traveler....Well...	Dozens of residents of Wuhan, the capital city...
2	eibebq	worldnews	2020-01-01 00:31:16	https://www.cnn.com/2019/12/31/politics/north-...	Kim Jong Un warns hostile US policy means ther...	10	This guy...Trump will comfort Kim....Being tha...	Washington CNN — \n\nNorth Korean Leader Kim Jo...
3	eibo17	worldnews	2020-01-01 00:53:25	https://www.theguardian.com/world/2019/dec/31/...	North Korean leader to end missile test ban, c...	7	[removed]...That's a choice of course. We can ...	Kim Jong-un has signalled that North Korea wil...
4	eibo8f	worldnews	2020-01-01 00:53:55	https://www.independent.co.uk/life-style/gadge...	Snapchat has stopped working for users around ...	8	Oh The Humanity ! ! !...OMG....what will teena...	For free real time breaking news alerts sent s...

Figure 11: Sample Reddit dataset

6.6 Data Cleaning

Cleaning posts for BERTopic: We are using a sentence transformer for embeddings that utilizes the structure of the sentence, so we performed light data cleaning without removing the stopwords and refrained from changing the grammatical structure of the sentence as they have an impact on the performance of the language model. We removed special characters, hyperlinks and extra whitespaces. A sample lightly cleaned post is displayed below:

Uncleaned post: https://www.theguardian.com/australia-news/bushfires
'Disaster's in the recovery': bushfire survivors still waiting for homes

Lightly cleaned post: disasters in the recovery bushfire survivors still waiting for homes

Figure 12: Clean Data Sample

Cleaning posts for Bag-of-words/TF-IDF LDA: To create a bag of words, we first lemmatize and then stem the text while excluding stopwords exceeding three characters in length.

```
Original document:
['SpaceX', 'successful', 'in', 'final', 'test', 'before', 'manned',
'flight', '—', 'NASA', 'said', 'it', 'was', 'the', "'last", "milestone'",
'before', '2', 'of', 'its', 'astronauts,', 'Doug', 'Hurley', 'and',
'Robert', 'Behnken,', 'are', 'ferried', 'to', 'the', 'ISS', 'with', 'a',
'SpaceX', 'instead', 'of', 'a', 'Russian', 'Soyuz', 'rocket.', 'SpaceX',
'has', 'said', 'it', 'hopes', 'to', 'launch', 'its', 'first', 'flight',
'for', 'tourists', 'in', '2021.']

Cleaned document:

['spacex', 'success', 'final', 'test', 'man', 'flight', 'nasa', 'say',
'mileston', 'astronaut', 'doug', 'hurley', 'robert', 'behnken', 'ferri',
'spacex', 'instead', 'russian', 'soyuz', 'rocket', 'spacex', 'say', 'hop',
'launch', 'flight', 'tourist']
```

Figure 13: Bag of Words Clean Data Sample

7 Conclusions and Discussion

The topic model and the accompanying set of visualizations developed by our team were successful in leveraging Reddit posts, user comments and linked content to streamline the experience of tracking a topic of interest and following how it evolves over time for the end user. Compared to other available tools, our implementation successfully incorporates a time-element to topic modeling. In addition, we look to leverage the advantages of sourcing content from a social platform by offering sentiment analysis through user comments to offer a holistic outlook on the top trending topics on Reddit, see how they relate to one another, and track how they evolve over time.

Future work should include generalizing the existing model to incorporate a large number of subreddits to truly gauge and keep track of how the social platform evolves over time. For example, some of the top topics in the "News" subreddit were Coronavirus, peace talks between Israel and Palestine, and the Paris protests in January 2020. In the same month, some of the top topics in the "Politics" subreddit were Coronavirus, Georgia senate elections, and Wisconsin election results. In the "Movies" subreddit, some of the top topics were Star Wars, Avengers, and movie trailers. In this way, various trending topics can be tracked each month using the visualization dashboard. Scaling-up also necessitates that we increase the time window used from roughly 2.5 years up to a maximum allowable value of 15 years. Going this far back in time when scraping linked content and news articles presents a problem as invalid URLs and non-existing pages start to become a problem.

Scraping user comments and other linked content tends to be very time consuming, therefore parallel processing and cloud computing emerge as possible novel solutions to speed up the data sourcing step. While our implementation is limited to Reddit, it is also possible to generalize the model used to process content across multiple social networks that allow content sharing and interaction between users. In terms of improvements regarding UI design, a feature of our model that has room for improvement is that the end user is currently able to see only three months at a time. While a user can use the next feature to identify topic trends in future months, it could be more beneficial to see changes in topic trends over a longer time frame such as a year.

7.1 Team Effort Distribution

Overall, throughout the length of this project, all team members have contributed a similar amount of effort.

References

[1] Rania Albalawi, Tet Hin Yeap, and Morad Benyoucef. Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3, 2020.

[2] Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *2008 Eighth IEEE International Conference on Data Mining*, pages 3–12, 2008.

[3] Animashree Anandkumar, Dean P. Foster, Daniel J. Hsu, Sham M. Kakade, and Yi-Kai Liu. Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. *CoRR*, abs/1204.6703, 2012.

[4] Claus Boye Asmussen and Charles Møller. Smart literature review: A practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1), 2019.

[5] Stephan A. Curiskis, Barry Drake, Thomas R. Osborn, and Paul J. Kennedy. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing Management*, 57(2):102034, 2020.

[6] Ali Daud. Using time topic modeling for semantics-based dynamic research interest finding. *Knowledge-Based Systems*, 26:154–163, 2012.

[7] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

[8] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, page 80–88, New York, NY, USA, 2010. Association for Computing Machinery.

[9] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014.

[10] Thomas Jacobs and Robin Tschötschel. Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, 22(5):469–485, 2019.

[11] Pooja Kherwa and Poonam Bansal. Semantic n-gram topic modeling. *EAI Endorsed Transactions on Scalable Information Systems*, 7(26):1–12, 2020.

[12] Yunhwan Kim and Sunmi Lee. #shoutyourabortion on instagram: Exploring the visual representation of hashtag movement and the public's responses. *SAGE Open*, 12(2):21582440221093327, 2022.

[13] David Mohaisen. *Computational data and Social Networks: 10th International Conference, csonet 2021, virtual event, November 15-17, 2021: Proceedings*. Springer, 2021.

[14] John W. Mohr and Petko Bogdanov. Introduction—topic models: What they are and why they matter. *Poetics*, 41(6):545–569, 2013. Topic Models and the Cultural Sciences.

[15] Eunhye Park, Woo-Hyuk Kim, and Sung-Bum Kim. What topics do members of the eating disorder online community discuss and empathize with? an application of big data analytics. *Healthcare*, 10(5):928, 2022.

[16] Florian Rabitz, Audronė Telešienė, and Eimantė Zolubienė. Topic modelling the news media representation of climate change. *Environmental Sociology*, 7(3):214–224, 2021.

[17] Jonas Rieger, Carsten Jentsch, and Jörg Rahnenführer. RollingLDA: An update algorithm of Latent Dirichlet Allocation to construct consistent time series from textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2337–2347, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[18] Kumar Shubankar, AdityaPratap Singh, and Vikram Pudi. A frequent keyword-set based algorithm for topic modeling and clustering of research papers. In *2011 3rd Conference on Data Mining and Optimization (DMO)*, pages 96–102, 2011.

[19] Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 424–433, New York, NY, USA, 2006. Association for Computing Machinery.

[20] Anke Wonneberger, Iina R. Hellsten, and Sandra H. J. Jacobs. Hashtag activism and the configuration of counterpublics: Dutch animal welfare debates on twitter. *Information, Communication & Society*, 24(12):1694–1711, 2021.

[21] Guixian Xu, Yueting Meng, Zhan Chen, Xiaoyu Qiu, Changzhi Wang, and Haishen Yao. Research on topic detection and tracking for online news texts. *IEEE Access*, 7:58407–58418, 2019.