Glance: Visualising Reddit with Topics Over Time

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

1 Introduction

People nowadays talk about various subjects on social media platforms every day to share their thoughts, find answers to their questions and debate. As more and more information is collected on the internet, it gets really confusing to follow a topic of interest. The objective of this project is to visualize different topics of discussions in Reddit posts and how those topics are evolving over time. Our goal is to build an interactive visualization tool for the users to find relevant topics and detect emerging topics or trends. It is hard to keep up with all the news and discussions pouring in all domains like Technology, Politics, Health, Sports, etc. It is even harder to translate that information into valuable insights. Our motive is to detect and keep track of the emerging topics and extract more useful and hidden information from Reddit for easily utilization by different organisations to plan ahead and make data-informed decisions.

2 Literature Survey

2.1 Topic Models

One of the popular topic modeling algorithms like LDA was used in conjunction with KNN to group information as well as identify if that information was fake or not. However, this approach did not perform well on conventional datasets despite categorizing the data into distinct categories [11]. It is also difficult to identify the right parameters in the model and optimize them [4]. Another widely used topic model is the Correlated Topic Model. It was also used to assess the relationships between discourse materials. CTM works 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 [14]. 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, better visualization could have been used to illustrate the relationship between different topics [13]. To perform semantics-based dynamic interest finding, a topic modeling approach is implemented which can implicitly capture word–word, word–author and author–author, word–time, and author–time relationships, simultaneously [7].

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 not capture time dependence of word occurrences [19]. Controlling topics in a way that makes it constant will deteriorate the performance of the topic model and produce unnecessary topics [17]. Online LDA has a beneficial effect on small subsets of data as it is efficient in both memory and speed [2]. Not only the topics but the actors of the topics are essential to consider while building a topic model or a lot of contextual information will be lost [8]. Utilizing the update algorithm along with LDA updates the models in real time without needing to recalculate the whole model, 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 [15]. 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 of the data [12][9]. Nevertheless, topic modeling must evolve to adapt for different document lengths [6].

2.2 Clustering Algorithms

K-Means is one of the most popular clustering techniques in literature and industry. This model was used to group both visual and textual documents that are alike which brings in additional dimensionality to the model to help improve its decisions. Yet, the number of clusters must be mentioned in advance to the algorithm which cannot be known earlier [10].

2.3 Other Methods

Newman modularity algorithm is a hierarchical algorithm implemented to measure the degree of interconnection inside a group and intraconnection between groups. It is instrumental for dimensionality reduction. However, it is not deterministic and can lead to different groups per run [18]. Excess Correlation Analysis is another type of 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. But it is unstable as it needs randomization to separate the singular values when computing Singular Value Decomposition [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, overlapping clusters using the topic as a similarity measure [16].

3 Innovation

In our project we are building a tool that implements a topic model on reddit posts while also taking into account the time component of the posts. There are topic models that has been implemented on reddit datasets but we have not found any approach which also deals with how a topic is evolving over time for the reddit posts. Also, we have not found an effective visualisation tool that detects emerging topics, trends, topic development and timeline of topics for reddit posts in one place. Our tool would be a one stop shop solution for effectively translating the reddit posts into valuable insights.

If successful we would be able to effectively organise and represent huge amounts of unstructured data, discover emerg-

ing topics in reddit, identify the hot and trending topics and visualise how a topic is evolving over time. This could provide valuable insights into the current market, what people are discussing and when a particular trend or event first started. For example, the GameStop short squeeze is primarily triggered by the subreddit r/wallstreetbets resulting in major financial consequences. With this tool, financial firms/institutes and regular traders can keep track of the emerging topics on social media efficiently and plan ahead.

4 Evaluation and Analysis

We can measure the success of our tool based on adaptation of the tool (number of page visits), user feedback on its ease of use and intuitiveness and the performance of our model. For model evaluation we will be employing both the observation based evaluation and quantitative metrics. In observation based evaluation we will look at the top words coming up in the topic and see if those makes sense. We will also interpret the topics based on the given words to see if the model is giving us coherent topics. In terms of quantitative metrics we will look at the perplexity scores, silhouette scores, KL divergence etc for the goodness of our model. We will also take advantage of having subreddits as the ground truth to see if the detected topics are falling under those subreddits.

5 Risks and Payoffs

In our project we are dealing with reddit posts which are unstructured and very unclean and will require a lot of data preprocessing. Also, the data source is humongous which could lead to huge storage and compute requirements. On the other hand, we would be building a tool to organise such huge amount of data effectively, discover hidden themes and hot trends in the data. The document representations based on topic distribution could be helpful in downstream tasks like recommendation, search, classification etc.

6 Tech stack and Cost Estimation

For tool development, we are planning on using Pushshift API for data collection [5], Pyspark for data cleaning and preprocessing, PostgreSql for data querying, Django, JavaScript (jquery, D3), HTML/CSS for web development, local drives for data storage. Initially, we are expecting that it will cost nothing for us to develop and implement the model for a smaller-sized subreddit. Later on when we scale and include numerous subreddits and decide to offer updates in real-time, the complexity and scale of computations will be increasing rather quickly. We could look to use cloud computing services that could either be free such as Google Colab, or paid services such as Microsoft Azure, Amazon AWS and Google Cloud within a free trial period. After that, the costs will heavily depend on the number of users and the size of the data being transferred and operated on. Some rough estimates: For storing the corpus on AWS s3 (Storage), it is $0.021 per GB. AWS Athena(Query) $5 per TB data scanned and $2/hr AWS EC2 (48 vCPU, 192 GiB Memory) instance for compute, Amazon Light sail for hosting $20 per month (4 gb memory, 2 core processor, 80 GB, 4 TB transfer).

7 Project Timeline and Milestones

For a group comprising six members, we are expecting the project to be completed in one-and-a-half to two months and will be dedicating at least 6-8 hours of time per week per member in the first few stages which translates to 200-300 hours of work in total.

Our milestones are divided into three phases as shown in figure 1. Our midterm would be the steps completed till Model Planning and Coding i.e. data collection, cleaning and preprocessing and building the topic model. After the midterm we would focus on building the visualisation, improving and adding the functionalities. Final exam would be finished tool that helps visualise the topics being discussed in reddit posts and their development over time. Meeting deadlines along with preparing and presenting the required deliverables could also be considered as additional benchmarks of progress and success.

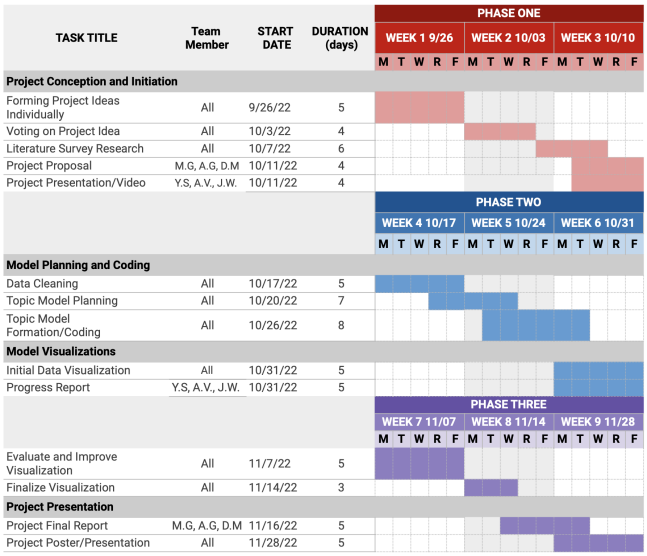

Figure 1: Gantt Chart

For the proposal tasks, all members contributed to the literature survey while Mehmet, Asha, and Dipendra worked on the proposal, Anshit and Jaylen worked on the presentation slides and Yashwant is presenting the video. All other non-report/non-presentation tasks will be evenly distributed between all group members who will each give an equal amount of effort.

8 Heilmeier Questions

1. Q1: Section 1
2. Q2: Section 2
3. Q3-Q5: Section 3, 4
4. Q6: Section 5
5. Q7: Section 6
6. Q8-Q9: Section 7

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. Online 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] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. *CoRR*, abs/2001.08435, 2020.

[6] 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.

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

[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] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.