GLANCE

Visualizing Reddit with Topics Over Time

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

Motivation/Introduction:

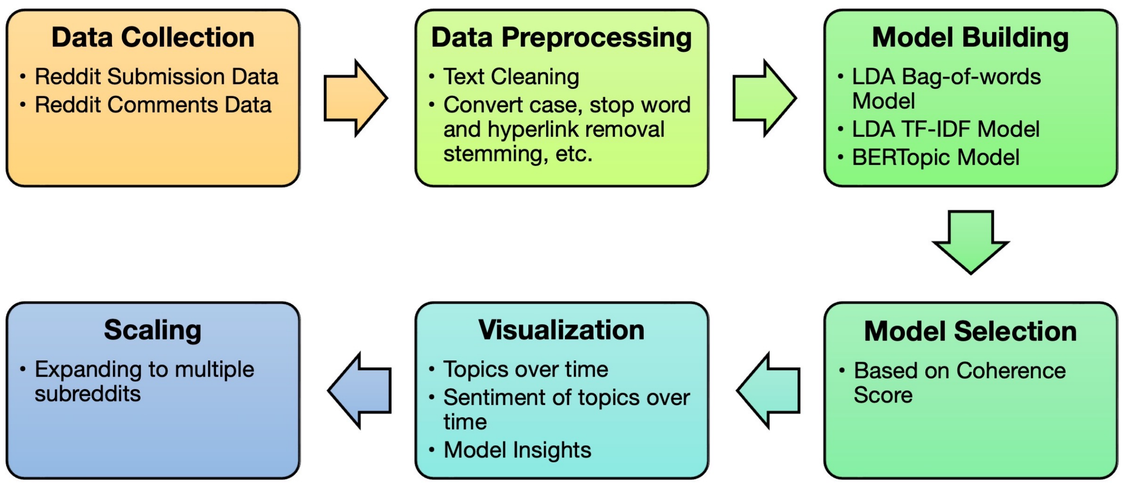
As more and more information is collected on the internet, topics of interest can be difficult to follow. It is hard to keep up with all the news and discussions pouring in all domains like Technology, Politics, Health, Sports, etc. It is even harder to translate that information into valuable insights. Our motive is to detect and keep track of emerging topics and extract more useful and hidden information from Reddit for easy utilization by different organizations to plan and make data informed decisions. The objective of this project is to visualize different topics of discussion in Reddit posts/comments and how those topics evolve over time. Our goal is to build an interactive visualization tool for the users to find relevant topics and detect emerging topics or trends.

Dataset:

We used the Pushshift API for Reddit to scrape posts, comments and metadata across three subreddits (r/worldnews, r/movies and r/politics) submitted between January 2020 and September 2022. Our dataset comprises over 300,000 posts and 2,250,000 comments.

Characteristics		
> 300,000 posts	1.46 GB	Temporal

Approach and User Interface:



Our approach uses topic modelling to extract high frequency topics on Reddit. This approach is successful as it allows users to see what kinds of topics remain relevant the longest and if certain times of the year have similar topic popularity trends. In addition to introducing a time component to topic modeling, other innovations of our approach include embedding linked content sourced from associated URLs and user comments within the corpus and using sentiment analysis to analyze user response and capture the relationship between user response and the relevancy of trends.
Glance uses BERTopic, a topic modeling algorithm that uses transformers and TF-IDF (word importance and frequency) to generate topic representations.

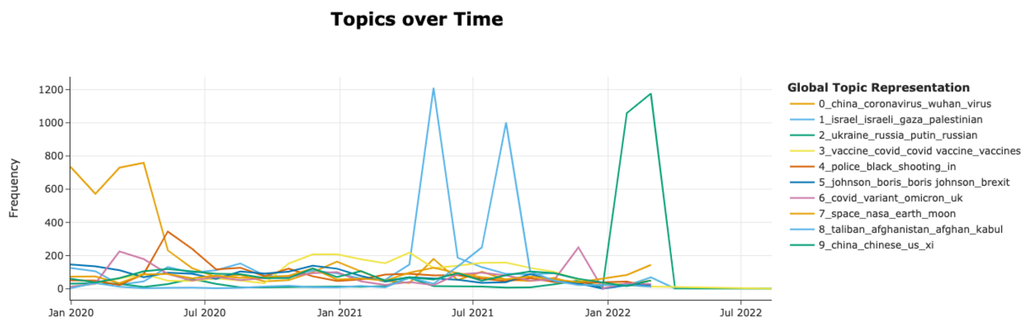
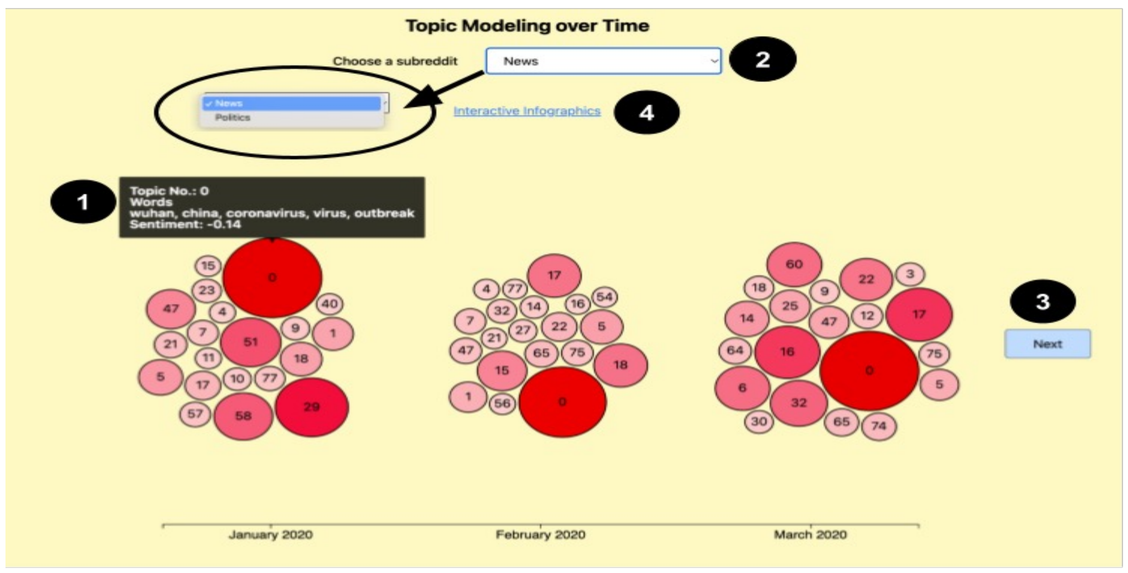


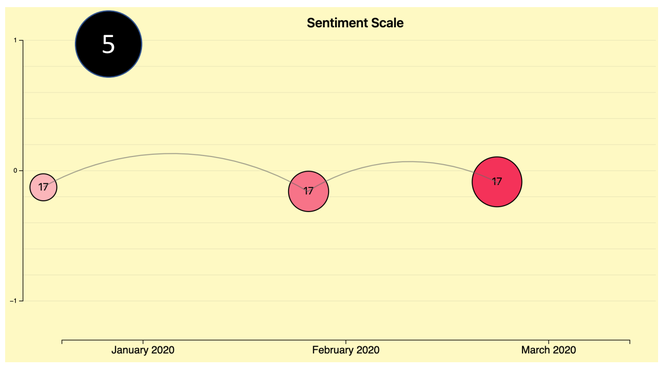
Figure 1: Topic Frequency Trends over time: helps the user determine the prevalence/popularity of topic at any given time and track how it changes.

Summary

Glance is an analytics tool that combines **visualization**, **user interaction**, **statistics**, and **topic modeling** to assist a user's exploration of important and trending topics on Reddit in a given time period.



The Glance user interface displays the top 20 topics for each month on Reddit data, where darker and larger topic bubbles represent the most frequent topics mentioned. **1**: When hovering over a topic one can find all the relevant words of the topic and the sentiment value. **2**: A user can filter for different subreddits, i.e. news or politics. **3**: The interface allows for topic. discovery over time by spanning a timeline of data. **4**: For more information on the BERTopic results, the link provides more interactive visualization (i.e. Figure 1) on the topic modelling algorithm the back-end. **5**: Once a topic is clicked, the user is brought to the screen on the left that shows how a topic's importance and sentiment is evolving over time.



Experiments and Results:

Prior to choosing BERTopic, we evaluated the method alongside Bag of Words with LDA and TF-IDF with LDA, assessing performance of each using coherence scores. BERTopic resulted in the highest coherence score of the three methods.

Method Options	LDA w/ Bag of Words	LDA w/ TF-IDF	BERTopic
Coherence Score	0.3	0.37	0.49

While there are other visualization approaches for topic modeling, Glance differs by showing top 20 topics, visualizing a smaller set, allowing for the user to easily identify topic trends without cluttered information. The algorithm in Glance is also scalable to visualize a larger set of top topics if necessary and scalable to incorporate more subreddit data.