Project Progress Report

Sentiment Analysis on Predicting Presidential Election: Case Study on Reddit Posts Using Firebase Realtime Database

Team 67

Armand Patel, Joshua Payapulli, Paul Yoo

DSCI551 Foundation of Data Management
Prof. Wenseung Wu

Sentiment Analysis on Predicting Presidential Election: Case Study on Reddit Posts Using Firebase Realtime Database

1. The Team Details

> Database, Data Distribution and Scaling Approach

(1) Database:

- Google Firebase

Data collected from Reddit is stored in four different Firebase real-time databases, three databases for distributed data storing, and one for raw data storing to back up all collected data in case of unexpected data loss.

Google Firebase database has advantages when implementing the project as it 1) provides data storage for free, 2) is capable of storing large amounts of data, 3) is faster in application development, 4) allows safe storage based on NoSQL, and 5) could share data easily among project members.

(2) Data Distributed Scaling Approach:

- Partitioning, Replication, and Horizontal Scaling

Databases are essential for the efficient functioning of any system relying on information retrieval and storage. When a database is overwhelmed by requests or lacks storage space, system performance may suffer, resulting in sluggish response times. Therefore, it's vital to address database scalability to meet the expanding demands for data storage and performance of a system. Approaches for Distributed scaling of data are to increase the database scalability.

There are a few database scaling and distributing approaches. 1) Horizontal scaling or Vertical scaling, 2) Replication or Sharding, 3) SQL or NoSQL Database, and 4) Tools and techniques, including load balancers, caching, and monitoring.

For the data distribution and scaling, the team applied three different approaches to enhance the project's success. 1) Data storage balancing through partitioning in three separate real-time databases using a hash function, 2) data replication and storing in a separate database for possible loss or damage in collected data, and 3) horizontal scaling by using three different databases to reduce computation time in future analysis processes.

2. Planned Implementation

> Plans and Changes from the initial proposal

We have had a few deviations from our original project proposal. We decided against the pulling of tweets from Twitter (X) API due to an updated policy to commercialize the developer tools where X no longer supports free data collection on tweets from other users. This monthly payment and limits on the amount of data that can be retrieved (1500 tweets per month) led us to decide to use Reddit as our alternative source of data. Reddit allows direct scraping of data from their website, and the use of an API to pull information. We've developed a file titled pull_data.py to pull Reddit content from the subreddit "r/Politics", which pertains to the topic of our project.

The Reddit API is called PRAW and can easily pull Reddit posts with a simple keyword search. The firebase DB allows for appropriate scaling. Currently, the team has collected 389 data rows from Reddit that include title, url, date of the post, and submission score which highlights the relevance of that post (upvotes/downvotes). We will be collecting more data while processing data for cleaning.

3. Status of the project

> Completed and Planned Milestones

As planned in the initial project proposal, the team has completed setting up an interactive workspace for project members to have effective communication during the project. Using the GitHub repository system, a safe and easy development environment has been created. The team has in-person or online meetings every week to check the progress of the project.

We have implemented Firebase real-time databases and have written code to pull data from Reddit using an API. Developed functions using a simple hash function to distribute rows of data across our databases with approximately equal load balancing. After sample data collection, we moved to real data collection based on approaches for data distribution and scaling. Recently, we've started data pre-processing as it has been planned from the beginning of the project.

In the future, we are planning to do sentiment analysis on our collection of Reddit posts to try predict the result of the US presidential election in 2024. While working on the analysis, web-based application development will proceed upon the completion of data pre-processing to prepare the Demo & implementation day by April 15, 2024.

4. Challenges

> Commercialized Tools, Analysis, Lack of experience in Firebase

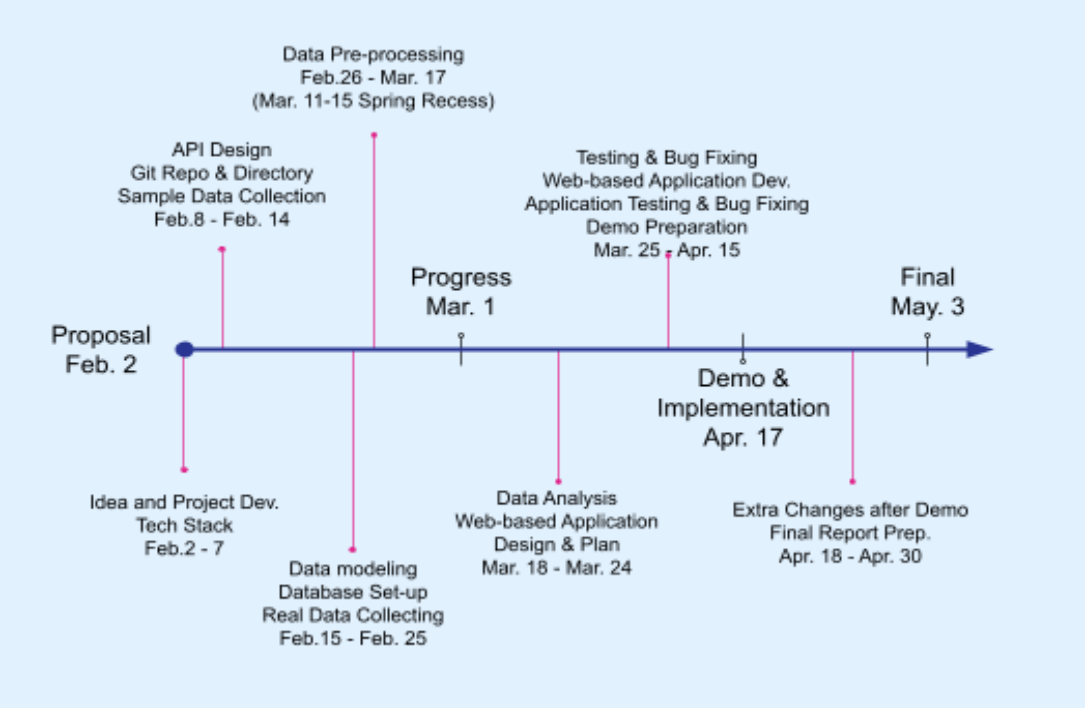
As a student working on the project, the greatest challenge is finding free tools and databases that could be utilized for the project. All members are willing to pay for the project expenses, but as students, we have decided to challenge ourselves with the harsh environment to strengthen our ability to solve problems.

While working on data preprocessing before analysis, we've found that we might need more attributes/fields to consider for better prediction. However, we have not decided which attributes to add to bring up the performance of prediction.

In addition to the two challenges, we've encountered challenges in handling the Firebase real-time system. Even though Google provides the dashboards and monitoring system for the database, it is only provided to subscribed members. We need to figure out how to handle and check the status of databases through programming languages while acquiring knowledge of guidelines provided by Google.

5. Timeline

> Milestones and Timeline



Until today, the team has successfully followed the initial plan. We will keep tracking the first project plan we've set and will be prepared for Demo & Implementation Day on April 17th.

By next week, the team will complete the data pre-processing, and start working on the Data Analysis and web-based application development process. The expected first testing date of the developed application is the end of March. Refer to the timeline table provided.

6. References

> Distributed Scaling of Data

Definition of Data Distribution and Scaling

- https://www.codecademy.com/article/database-scaling-strategies

Data Distribution and Scaling Approaches

- https://www.linkedin.com/advice/0/what-factors-should-you-consider-when-scaling-8f5me#:~:text=Partitioning%20and%20replication%20are%20two,assigning%20them%20to%20different%20nodes

Horizontal scaling

- https://www.linkedin.com/advice/0/how-do-you-scale-distribute-your-database#:~:text=Scaling%20and%20distributing%20a%20database,resolution%2C%20and%20monitoring%20are%20key