

Project Status Report: Stock Recommender System

[Team Members] [Project Overview] [Project Status] [Project Challenges] [Proposed resolution] [Next Steps]

Team Members

Name	NetID	Email ID
Ezra Schroeder	ezras2	ezras2@illinois.edu
Rasbihari Pal	pal9	pal9@illinois.edu
Sandeep Kumar	kumar64	kumar64@illinois.edu

Team captain marked in **BOLD**.

Project Overview

We are building a product "Stock recommender system", which will enable retail investors to make informed decision about their investment choices, based on market analysts ratings and market/investors sentiments at the moment. Based on user's preference of a stock, the product recommends 5 stocks which closely matches the user's preferred stock and ratings for them.

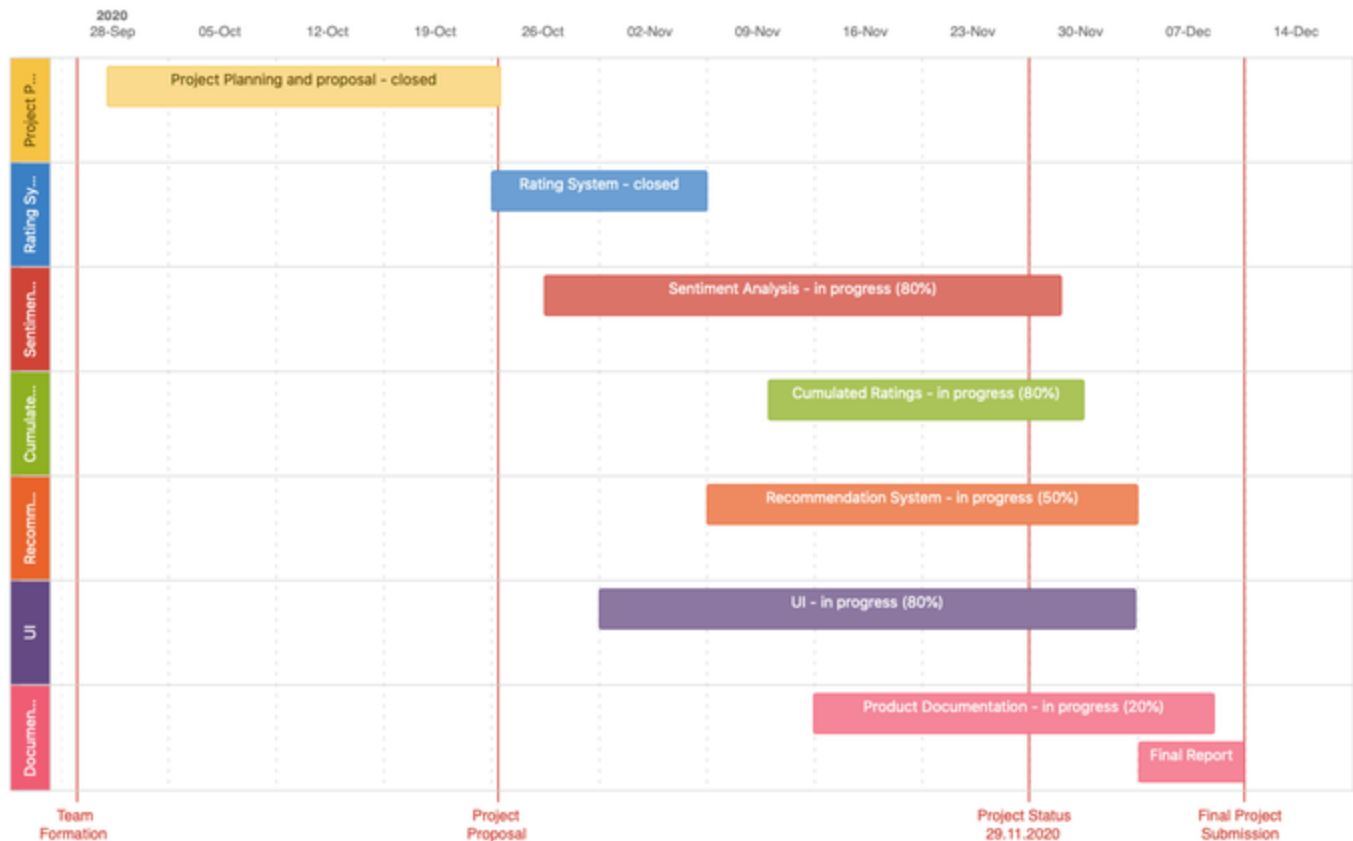
The recommender system combines

- rating data from various market analyst,
- market sentiments using microblogging data (using twitter) and
- company profile

for determination of the overall rating and curation of the recommendation list.

Project Status

The team has been making steadfast progress in various topics in order to deliver the project on time by the final project submission deadline. Progress on each individual topic can be seen in below actualized project planning chart.



Percentage completion of the tasks under individual topics is mentioned in brackets () next to each topic name in above planning chart.

Project Challenges

At the conception of the project, we had foreseen some challenges but during the execution, we faced several challenges, which we had either not foreseen or assumed as minor challenges. Below, we list of the challenges which required significant time, effort and energy from all of us.

- Working with Twitter API and gathering training and test data for the model we prepared for sentiment analysis
- Collecting good quality financial data of various companies
- Scraping of market analysts websites
- Technical limitation to process the data

Above challenges are in addition to a major hurdle related to working in a team with members sitting across Atlantic Ocean and having a time difference of 7 hours. Thanks to team members flexibility, team agility and technology, this challenge did not pose a big threat to our continuous progress.

Proposed resolution

We, as a team, invested significant amount of time working on above listed challenges. Some of the challenges we could overcome with sustained effort but for some of the challenges we had to find a middle ground to meet the project objective of delivering a 'reasonably' working solution and hoping to improvise the solution over time with larger effort.

- Working with Twitter API and gather training and test data for the classifier

It required significant effort and multiple round of communication with Twitter team to get access to the Twitter API - an important part of our proposed solution for performing the market sentiment analysis for a given stock. Eventually we received the access but with restriction on the volume of tweets we can pull in a month.

For training and test data, we used public data available in internet. We used the following polarity data set:

<https://www.cs.cornell.edu/people/pabo/movie-review-data/>

This helped us to avoid the significant manual effort for downloading the twitter data and possibly human labeled them.

- Collecting good quality financial data of various companies

For building a good recommendation engine, access to financial data of various companies is required. We could not find a single source with latest financial data for the listed companies. To avoid the trap of spending significant time and energy to resolve the issue, we eventually settled with a kaggle data source having financial data from 2018. But the solution is being designed in a way that we can update the data any time to reflect current reality.

- Scraping of market analysts websites

Our initial idea was to scrape websites and reports from various market analysts for their ratings to various stocks. After collating the information for couple of stocks, we realized the vastness of this topic. For each stock, we needed to scrape data from various market analysts and each report following free text structure. For our project, we needed access to the assigned rating from various market analysts and we settlement with an established website, which collects and publishes such ratings in structured format.

- Technical limitation to process the data

At the moment we have collected financial data for more than 1800 companies. To scrape the analyst ratings and get twitter feed for all these companies and then processing them for final result would require significant time and processing infrastructure. As a realistic approach, we have decided to limit the scope of the project to a limited list of stocks. This limit is only for practical purpose only. However, the solution is capable to manage more volume of data with parallel and batch processing.

Next Steps

At the moment, we do not see any major technical challenge in our way in coming days. We are planning to complete the remaining tasks from all topics and focus on final project delivery on time.