Reaction Paper(-0.5): The survey paper was great but you could have talked more about the key takeaways rather than the questions it is addressing. **DONE**

Completion(-0.5): To measure success, you can also think about certain baseline systems and compare with them.

Quality(-1.5): While there is some level of novelty in your hybrid solution, you can also try to incorporate some of the ideas you had in the brainstorming sections (e.g. incorporating things like amount of time spent by a user on the news or maybe doing online news recommendation). **DONE**

Here is the feedback on your milestone report.

Overall, I was really happy with the progress you made and with the fact that you were able to run a basic version of LDA and see that you can only get high-level topics from it. This is where you lost points:

Scientific Merit and Technical Depth: -1.5 (There were still no plans for baseline, I would have also liked to see a basic description of how hLDA works)

Presentation: -1 (Typos, structure, references)

Going forward, I have the following suggestions:

(1) Please have 2-3 (maybe 4) baselines and 2-3 ablated versions of your model (e.g., without user interest, without smoothing factor, etc.) and compare against them. This would be really helpful in analyzing what the benefits of using your model are and where the gains of your model are coming from.

(2) Be a little more structured in your final report and follow best practices in academic writing. I would recommend using the standard referencing format (\cite{} in latex) and avoid using references as a noun (e.g., "(2) shows that..." , say "Author et al. [2] shows that.."). Try saving your images as ".pdf" instead of ".jpg" or ".png". You will notice a huge change in the quality of rendered images. Also, some portions such as Exploratory Data Analysis could have been more modular with better structure and motivations (e.g. answering why are you looking there, why are you doing a particular experiment). In fact, you should try to answer the "why" question wherever possible (e.g., in the various components of your model). While deducting points on this may sound harsh, we really care about various things you can learn while working on this project.

**Abstract**

Studies have shown that social networks and search engines are associated with an increase in an individual’s exposure to news content. While there exists a lot of research experimenting with various technical approaches to personalize news recommendation for readers, news recommendations on social media platforms is an advancing field. The objective of our project is to build a news recommender system by analyzing user interactions in social networks and combining those with results of text mining approaches to generate personalized results. The motivation behind this project came from the literature review that shows a lack of user-side contextualization in news recommendation. Among the different domains of recommender systems, news recommendation has been explored relatively less due to the lack of structured data and features. Text mining for news articles using NLP techniques in itself is a different class of problem. In this project we have built a pipeline for extracting user and news article features, to create a hybrid recommender system that addresses the problems of cold start, data sparsity and scalability.

**Introduction**

Today the world is flooded and overwhelmed with news. According to our literature review, more than 2 million articles are published every day on the web. The online consumption of news is further fueled by their dissemination on various online social platforms. According to studies done at Pew Research Center, 55% of U.S. adults now get their news from social media either "often" or "sometimes" – an 8% increase from last year.

This explosion of online published content resulted in various researches on the personalization of news feed for users. Our literature surveys show that most of this research is driven by content-based recommendation techniques, which means recommending users articles based on their viewing history. We realized that this approach lacks a critical element of inculcating user-side interactions with the news and soon realized that the problem lies with the limited user-interactions indicators, click rate and time spent seems to be the most common.

We have tried to mitigate this problem by analyzing user networks on social media platforms. While harvesting metrics such as click rates and time-spent on reading a news article can be difficult, free APIs can be used to gather information on social networks. We have used Twitter data for our project.

In this report, we propose a method for news recommendation on social media platforms. Our recommendation engine employs user side information as well as the news article information for a hybrid Collaborative Filtering and Content-Based Filtering(CF-CBF) approach, to recommend a news article to the user.

We have experimented with both vectorised and graph based approaches to cluster our users based on their news consumption patterns. Our method uses hierarchical clustering to group the users into clusters based on the number of retweets. On the item side clustering, we again experimented with vectorised approach and probabilistic approaches to find article clusters, which also forms the base of our Content Based Filtering (CBF) score. Combining time trends and user interests we find the time specific user interest for each topic, thus getting a Collaborative Filtering (CF) recommendation score. Finally, we take the two recommendation scores and combine them to a single score and recommend top news articles based on the choice of the individual and topic of the article.

\section{Related Work}

In the domain of news recommender systems, there are 3 classes of recommendation algorithms- Collaborative Filtering (CF), Content-based Filtering (CBF) and hybrid algorithms. The paper \cite{ref1}, proposes the use of building a hybrid recommendation algorithm. They break down the recommendation engine problem into two phases - modelling and recommendation phase. The objective of the modelling phase is to pre-learn user and content models that serve to reduce the dimensionality as well as group similar users/articles together. These modellings are performed offline and stored prior to the recommendation phase. The recommendation phase occurs in real-time where based on a reader query, the results of the CF and CBF approaches are combined together to produce a set of recommendations. This paper was a critical read and helped understand not only the various recommendation techniques but also introduced us to the many evaluations and optimization techniques that are used to build recommendation engines\\

In paper \cite{CBFpaper}, the authors propose an interesting approach to calculate the recommendation score for the content filtering algorithm. The authors propose to compute the user’s interest in a category (or topic) based on the user’s overall interest in a topic and also the trend of interest for a topic over time. The main takeaway of this paper was that by combining these “long-term” and “short-term” effects, we can estimate the current interest of a user for a topic. In this project we will use a similar approach for the CBF recommendation. \\

The novelty of our approach lies in the network analysis for user-side clustering. The paper \cite{userClustering} introduced us to the approach of using retweet information to cluster tweets instead of tweet text. The justification for the approach comes from the fact that the tweets are usually restricted to 280 characters and words are often abbreviated (modern text language) and contain special symbols like hashtags and emojis. Thus the paper proposed a method to calculate the similarity between tweets based on the overlap rate of users who retweeted them, creating a retweet network followed by clustering the tweets based on network clustering. In this project we decided to use a similar approach to cluster the users based on a retweet network. This user modelling will be used in the CF recommendation approach. \\

Eventually to combine the recommendations, we will explore the use of a weighted additive recommendation score or a multiplicative score as proposed in \cite{CBFpaper}.