Feedback Based News Recommendation System

Samiksha Garg (MT21074)
samiksha21074@iiitd.ac.in
IIIT Delhi

Debnath Kundu (MT22026)
debnath22026@iiitd.ac.in
IIIT Delhi

Shambhavi Pathak (MT22067)
shambhavi22067@iiitd.ac.in
IIIT Delhi

Snehal Buldeo (MT22074)
snehal22074@iiitd.ac.in
IIIT Delhi

ACM Reference Format:
Samiksha Garg (MT21074), Debnath Kundu (MT22026), Shambhavi Pathak (MT22067), and Snehal Buldeo (MT22074). 2023. Feedback Based News Recommendation System. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM DEFINITION

The method we discuss in this paper, content diversification in news recommendation, aims to balance and diversify personalized recommendation lists to capture the user's broad spectrum of interests precisely. Feedback based refers to the process of explicitly asking the users' feedback before recommending a diverse list.

Thus, Introducing diversification in the news recommendation list only if the user wants to diversify its feed. As a result, people may only be exposed to content that reinforces their existing beliefs and interests, rather than being exposed to a wide range of perspectives. Filter bubbles can lead to a distorted view of reality, as people may be unaware of important issues and events outside of their bubble.

2 MOTIVATION

The reason behind diversifying the recommendation list is to avoid the over-personalization of the user's news feed and prevent filter bubbles or echo chambers. Filter bubble, alternatively called an echo chamber, refers to the way that algorithms on news recommendation systems, social media, and search engines can tailor the information that users see based on their previous search history and engagement patterns. As a result, people may only be exposed to content that reinforces their existing beliefs and interests rather than being exposed to a wide range of perspectives. Filter bubbles can lead to a distorted view of reality, as people may be unaware of important issues and events outside of their bubble. Our method improves diversity with recommendation lists, particularly those generated using the common Tf-idf, Bag-of-Words algorithm. Our work builds upon prior research on recommender systems, looking at properties of recommendation lists as entities in their own right rather than specifically focusing on the accuracy of individual recommendations.

3 LITERATURE REVIEW

The term 'user experience' may have different interpretations in a recommendation domain , such as usability, usefulness, effectiveness or satisfactory interaction with the system. The task of recommending appropriate and relevant news stories to news readers is challenging. The reason is that the news domain is faced with certain challenges that are different from those of other application domains of recommender systems.

However, overly personalized news stories limit readers' exposure to different types of news. At the individual level, a news reader may get bored of reading similar types of news stories all the time. Over-personalization may also affect a reader's behavior in the long run, causing them to avoid counter-attitudinal (attitude that contradicts one's own beliefs) information (viewpoints, opinions) . This type of behavior, at the societal level, poses a threat to democracy in the form of people's denial of opposing viewpoints.

Too much personalization in an NRS is often the result of recommendation approaches that place too much emphasis on prediction accuracy. These typical accuracy-centric approaches may fail to consider other aspects of subjective user experiences (such as choice satisfaction, perceived system effectiveness, better recommendations, and exposure to different points of view) when evaluating the recommendation quality. When developing a good NRS, one must consider the beyond-accuracy aspects to evaluate the quality of news recommendations

On the other hand, the diversity in news domain is crucial not only to keep readers engaged during the online reading process but also to expose readers to counter-attitudinal behavior [3].

Diversity measures the degree of 'dissimilarity' among the recommended items. It is mostly implemented through re-ranking of the recommendation lists. Some well-known metrics are: Intra-List Diversity (ILD) (diversity between any two items of recommended list). The traditional pairwise diversity ILD remains a popular metric to evaluate diversity in NRS. The ILD can be computed among the items, topics, categories, tags or even sentiments (tone) [2] in an NRS. Since the typical ILD method is computed for each individual user, it is a computationally expensive process for an NRS where there are millions of users and items. Thus, it requires more research to consider various aspects, such as level of diversification, scalability issues in an NRS.

4 NOVELTY

In order to accurately capture the user's wide range of interests, the method that is addressed in this paper, content diversification in news recommendation, balances and diversifies personalized recommendation lists. We present two novel ideas:

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- The balance between diversification and personalized recommendation, that is, making sure the recommendation is neither over-personalized with every iteration nor the recommendation is too different from the user's personal choice
- The diversification we added to the recommendation is purely content-based; it does not use collaborative filtering techniques to consider other users' reading patterns.

5 METHODOLOGY

The work in the project has two major components in it, the initial recommendation of the user, based on it's clickstream data and the diversification algorithm to diverse the recommendation before the user is subjected to over-personalization.

The clickstream data is collected to learn about the user's personal choice. Based on pairwise cosine similarity calculated on tf-idf vector, the top two hundred similar articles are considered, out of which top ten are displayed to the user.

If for a recommendation list the ILD (*Intra-List Diversity*) value crosses the defined threshold, we ask for user's feedback if they will like to diverse their news recommendation, if agreed upon, the diversification algorithm is invoked, which shuffles the articles in the recommendation list to rank the bottom articles, from the initial list of two hundred articles to the top and showcase it to user.

The diversification is introduced in the data using semantic analysis of the news content. The news content was scraped from the source URL of the respective news, following which sentiment analysis was performed to extract the following features:

- Polarity: It is a float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement
- Subjectivity: It is a float which lies in the range of [0,1] where 0 means the article is factual and 1 is a personal opinion
- Sentiment: This feature has been extracted from the content of the article. Its values are **positive** (polarity value ranging from (0,1]), **negative** (polarity value ranging from [-1,0)) and **neutral** (polarity is 0)

6 DATABASE

Initially the dataset we considered was BBC news dataset. After conducting data analysis, we discovered that the news dataset did not contain a uniform distribution of articles in each category, so we dropped it. Also, several news articles' source urls were out-of-date because they were from 2017. The writing style of the news article had changed as a result of the articles' age.

The dataset we used for the current project is made up of 1170 articles that were exported using the Inshorts API and was gathered over the course of 21 days. It has the following features:

- headlines
- short description
- source url
- category

The data was collected using the Inshorts topic-based API, which made sure that there were a consistent number of articles in each category. The article categories we have worked with are Politics, Business, Sports, Technology, Entertainment, Education, Fashion.

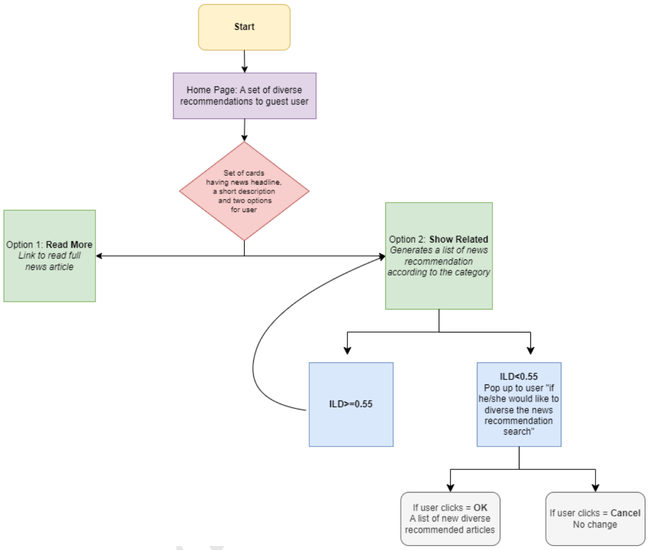
Start
Home Page: A set of diverse recommendations to guest user
Set of cards having news headline, a short description and two options for user
Option 1: Read More
Link to read full news article
Option 2: Show Related
Generates a list of news recommendation according to the category
ILD>=0.55
ILD<0.55
Pop up to user "if he/she would like to diverse the news recommendation search"
If user clicks = OK
A list of new diverse recommended articles
If user clicks = Cancel
No change

Figure 1: The flowchart of our website

6.1 Data Cleaning and Preprocessing

To create tf-idf vector over the original dataset, we needed to preprocess the data. As part of the data preprocessing we performed the following steps:

- Removing Duplicate articles
- Removing Punctuations
- Removing Stop-words
- Lemmatization
- Tokenization

We used modules from the nltk library for performing the above operations.

6.1.1 Topic Modelling. For the purpose of identifying the abstract topics that appear in a group of documents, topic modelling is used. To assign text in a document to a specific topic, **Latent Dirichlet Allocation (LDA)** , a type of topic model, is used. It develops a Dirichlet distribution-modeled topic per document and words per topic model.

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for topic modeling in natural language processing. It allows us to automatically discover hidden topics within a large collection of text data.

Using topic modelling a new feature was added to the dataset, **topic** which included 0, 1, 2, 3, 4, 5, 6, 7, 8 topics. Since LDA does not have any predefined topics and forms clusters of words which have a high probability of occurring together, thus the clusters were encoded as numbers.

6.1.2 Sentiment Analysis. As mentioned in 5 sentiment analysis was performed to get an extensive semantic understanding of the news content for content based diversification. To extract the complete article content we had to perform web scraping. For web scraping we used article module from the Newspaper3k library. As part of the web scraping an additional feature was added, where the content of the article was stored. Using the TextBlob module the sentiment analysis was performed as part of which three new columns were added namely polarity, sentiment and subjectivity.

7 EVALUATION METRIC

The recommender systems have been evaluated according to accuracy metrics that measure the algorithm performance by comparing its prediction against a known user rating of an item (Herlocker et al. 2004; Gunawardana and Shani 2009). However, such accuracy-centric evaluations cannot answer the question about if users are satisfied with the recommendations. For example, Amazon claimed to generate an additional 10% to 30% of its revenue in 2015 from the sale of diverse (non-personalized) items (Srihari 2015). This kind of insufficiency has shifted some researchers' focus to different goals for a recommender system, which can address other aspects beyond accuracy. Generally, recommending everything related to users' preferences would result in good accuracy. However, for news consumption, we discuss the beyond-accuracy aspect in NRS i.e. diversity

As mentioned in 3, diversity measures the degree of 'dissimilarity' among the recommended items. *Intra-List Diversity* is a popular metric to measure the diversity of a recommendation list. It can be calculated a follows:

$$\frac{1}{|S|(|S|-1)} \sum_{i_l \in S, l<k} d(i_k, i_l)$$

In the above formula, |S| refers to the number of samples in the recommendation list displayed to user i.e. 10 in our project.

$$d(i_k, i_l)$$

refers to the distance between any two articles in the dataset. The value of ILD varies between 0 and 1, where a value of 0 means that there is no diversity in the recommendation list and a value of 1 means the list is highly diverse. For our dataset, the value of ILD varied between 0.50 to 0.59 and after diversifying the list the value fluctuated between 0.67 to 0.76. The chosen threshold for showing a pop-up to user is 0.52, considering the range of the ILD value for the non-diverse recommendation.

8 CODE

Github repo url

9 EXPERIMENTAL RESULTS

We conducted an extensive range of experiments with various combination of news categories, different scores of the extracted features. Given below are our findings:

Figure 1 represents the impact of diversity factor on a recommendation list. The plot begins from the initial diversity score w.r.t diversity factor 0.0. As the factor is increased, the diversity also increases, thereby meaning that the news coverage is more diverse.

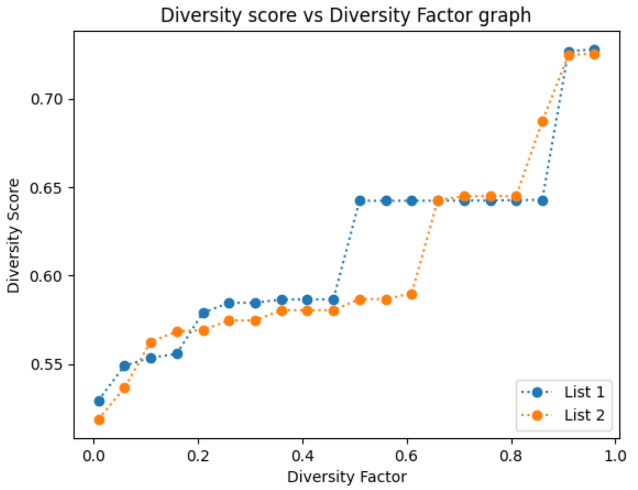

Figure 2: Each point is a diversified list at a particular diversification factor.

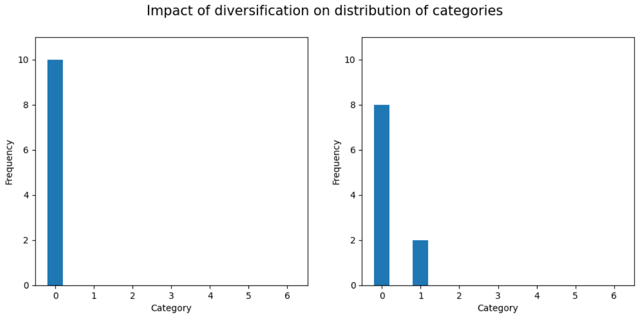

Figure 3: The main category distribution has improved.

Figure 2 clearly shows the impact of diversity factor on the **main categories** of the recommendation list. In the example shown, initially there were 10 news from the same category (label-encoded as 0). After applying the diversification algorithm, there were 2 recommendation from category 1 as well. So, the initial preference (i.e. category) of the user is preserved to some extent, while diversifying the content.

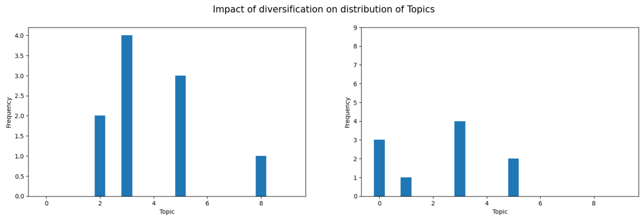

Figure 4: The sub-category distribution has a huge enhancement. User preference has been preserved to some extent.

Figure 3 depicts the distribution of sub-categories that were extracted by applying the **Latent Dirichlet Allocation(LDA)** topic modelling algorithm. Although the main user-preference is entirely preserved, however there are diverse recommendations based on the sub-categories extracted from the article text.

Figure 4 shows the **subjectivity** scores of every article in the original list of top 20 recommendations as well as the top 20 articles

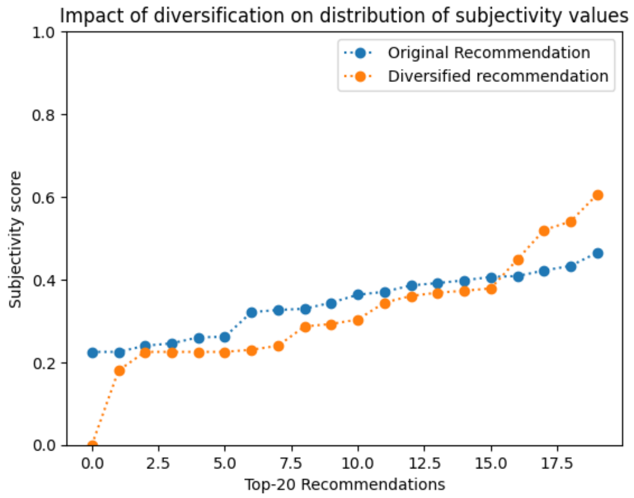

Figure 5: Subjectivity lies between [0,1]. The higher subjectivity means that the text contains personal opinion rather than factual information.

in the diversified list. The orange curve shows more range in its values, hence more diversified in nature.

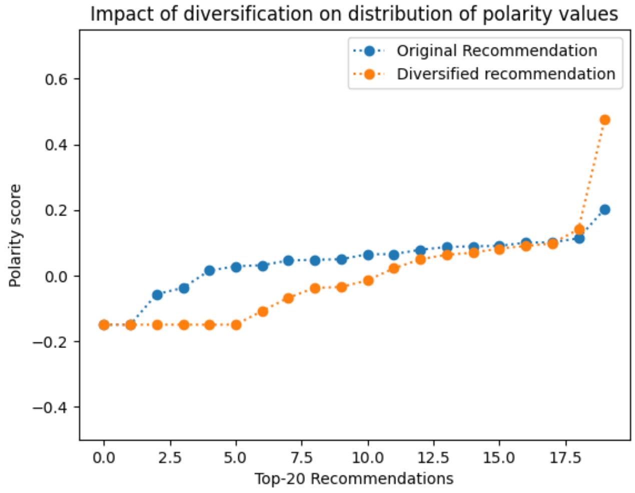

Figure 6: Polarity lies between [-1,1], -1 defines a negative sentiment and 1 defines a positive sentiment.

Figure 5 shows the **Polarity** scores of every article in the original list of top 20 recommendations as well as the top 20 articles in the diversified list. The orange curve shows more range in its values, hence more diversified in nature.

10 ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Prof. Rajiv Ratn Shah for his outstanding leadership and guidance throughout the project. His passion for the subject matter and tireless efforts have elevated the quality of our work to new heights. TAs for the course : Mr. Ritwik Mishra and Mr. Avinash Anand for their invaluable contributions to this project. Their exceptional technical expertise, in-depth knowledge, attention to detail, and problem-solving skills have been invaluable in overcoming challenges and achieving our project goals.

Mr. Sumit Kolhe (email: thesumitkolhe@gmail.com), author of GitHub repository of inShorts API for his quick response and resolution to the bug in API. Mr. Manvendra Kumar Nema (email: manvendra22038@gmail.com), our classmate has also helped us with his knowledge of Python and ML.

Member contributions: Debnath Kundu (MT22026) - Research, News article extraction(web scraping), Sentiment Analysis (polarity and subjectivity); Shambhavi Pathak (MT22067) - Research, Dataset generation (using InShorts API), Report ; Samiksha Garg (MT21074) - UI development, Integration ; Snehal Buldeo (MT22074) - Research, LDA (Topic extraction), Diversification of recommendations; Sakshi Sinha (MT22121) (till midsem) - Research, Literature Review

Last but not least, I would like to express my gratitude to our mentors, colleagues, friends, and family members for their encouragement, support, and motivation throughout this project. Their belief in us has been a driving force behind our success.

REFERENCES

[1] [n. d.]. BBC News Classification. https://kaggle.com/competitions/learn-ai-bbc.
[2] [n. d.]. TextBlob: Simplified Text Processing — TextBlob 0.16.0 Documentation. https://textblob.readthedocs.io/en/dev/.
[3] Wanrong Gu, Shoubin Dong, Zhizhao Zeng, and Jinchao He. 2014. An Effective News Recommendation Method for Microblog User. *The Scientific World Journal* 2014 (2014), 907515. https://doi.org/10.1155/2014/907515
[4] Natali Helberger. 2019. On the Democratic Role of News Recommenders. *Digital Journalism* 7, 8 (Sept. 2019), 993–1012. https://doi.org/10.1080/21670811.2019.1623700
[5] Sumit Kolhe. 2023. Inshorts News API.
[6] Joseph A. Konstan and John Riedl. 2012. Recommender Systems: From Algorithms to User Experience. *User Modeling and User-Adapted Interaction* 22, 1 (April 2012), 101–123. https://doi.org/10.1007/s11257-011-9112-x
[7] Lei Li and Tao Li. 2013. News Recommendation via Hypergraph Learning: Encapsulation of User Behavior and News Content. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13)*. Association for Computing Machinery, New York, NY, USA, 305–314. https://doi.org/10.1145/2433396.2433436
[8] Andrii Maksai, Florent Garcin, and Boi Faltings. 2015. Predicting Online Performance of News Recommender Systems Through Richer Evaluation Metrics. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15)*. Association for Computing Machinery, New York, NY, USA, 179–186. https://doi.org/10.1145/2792838.2800184
[9] Shaina Raza and Chen Ding. 2020. *A Regularized Model to Trade-off between Accuracy and Diversity in a News Recommender System.* 560 pages. https://doi.org/10.1109/BigData50022.2020.9378340
[10] Shaina Raza and Chen Ding. 2022. News Recommender System: A Review of Recent Progress, Challenges, and Opportunities. *Artificial Intelligence Review* 55, 1 (Jan. 2022), 749–800. https://doi.org/10.1007/s10462-021-10043-x
[11] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving Recommendation Lists through Topic Diversification. In *Proceedings of the 14th International Conference on World Wide Web - WWW '05*. ACM Press, Chiba, Japan, 22. https://doi.org/10.1145/1060745.1060754