TweetPress - Twitter News Article Recommender

Manas Agarwal
IIIT Delhi
manas20443@iiitd.ac.in

Kartik Jain
IIIT Delhi
kartik20440@iiitd.ac.in

Uttkarsh Singh
IIIT Delhi
uttkarsh20479@iiitd.ac.in

Rishabh Oberoi
IIIT Delhi
rishabh20459@iiitd.ac.in

Darsh Parikh
IIIT Delhi
darsh20560@iiitd.ac.in

23 March 2023

1 Introduction

1.1 Motivation and Problem Statement

Our project aims to solve the problem of people receiving news articles that they might not be interested in by creating a personalized news recommendation system. Our model takes the user ID as input and recommends incoming articles by using a similarity model based on the user's interests.

1.2 Novelty

The novelty of our project lies in creating a personalized news recommendation system that takes into account the user's interests based on their tweets and recommends relevant articles. For, the model implementation we use different information retrieval techniques. Most users use Twitter to get daily updates on what's going around and our model helps them to get personalised news articles that they might like or might be looking for.

1.3 Dataset Description

We scraped out a user's tweet using snscrape python library. For further use, the retrieved tweets were preprocessed, and hashtags and keywords were separated. For the news articles dataset, we used Newsapi.org and collected articles on different topics and for the mid-project, about 1800 news articles were collected. We have collected tweets from various famous personalities and tried recommending the news articles based on them.

2 Proposed Solution

We take TwitterID, Number of News articles and the Weighting Scheme from the user. After applying the models based on the weighting scheme mentioned by the user, we give the output as top X articles where X is equal to the number of news articles mentioned by the user. The workflow of our model is: Initially we extract the tweets of the user for the last 6 months. Then we make a corpus for the user's tweets. For the midsem review we were not able to make a dynamic model so we made different categories or tags for news articles like AI, Tech, Business, Climate etc. We used newsapi.org to extract the news articles and their categories. We treat a news article and its corresponding tag as a separate document and also associate it with a link. After creating these documents we make TF-IDF of all these documents based on the evaluation metric chosen by the user. Made a query vector using the corpus of the user's tweet made above. The size of the query vector will be the same as the TF-IDF matrix and it will contain the words that will be present in the intersection of the vocabulary of TF-IDF and the tweet corpus. For the midsem review, we have used the Cosine Similarity to find the similarity between the query vector and the news articles.

3 Literature Review

The paper proposes a news recommendation system for Twitter users that utilizes a hybrid filtering approach, combining collaborative filtering and content-based filtering techniques. The system uses Twitter user's tweet history and engagement data to identify the user's interests and preferences. It then recommends relevant news articles by considering the similarity between the user's tweet content and the news article's content. The paper includes an experiment conducted on a dataset of tweets and news articles to evaluate the proposed system's performance. The results showed that the hybrid filtering approach improved the recommendation accuracy compared to using either filtering technique alone. The paper also highlights the potential benefits of the system for news service providers to improve user engagement and retention.

The paper titled "A Review on Text Similarity Technique used in IR and its Application" provides an overview of text similarity techniques used in information retrieval (IR) and their applications. The paper discusses different approaches to measure text similarity, such as cosine similarity, Jaccard similarity, and Euclidean distance. It also highlights the importance of text preprocessing techniques in improving the accuracy of text similarity measurements.

The authors discuss various applications of text similarity techniques in IR, such as document clustering, topic modeling, and recommender systems. The paper focuses on the application of text similarity techniques in building recommendation systems, including content-based filtering and collaborative filtering. The authors compare and contrast the strengths and weaknesses of these techniques and highlight the need for a hybrid approach that combines both techniques to improve the recommendation accuracy.

Overall, the paper provides a comprehensive review of text similarity techniques used in IR and their applications in building recommendation systems. The authors conclude by emphasizing the importance of selecting an appropriate text similarity technique based on the specific application and the need to continually explore new techniques to improve recommendation accuracy.

3.1 Similar Work

The paper titled "Discovering significant news sources in Twitter" proposes a method to discover significant news sources on Twitter. The authors highlight the importance of identifying reliable and trustworthy news sources on Twitter, given the abundance of information available on the platform. They have made a framework which downloads a list of news-related relevant tweets from twitter, extracts URLs associated with those tweets and infers the significance of those URLs in Twitter. They collected tweets using the REST and Streaming API of Twitter, and then extracted URLs from these tweets. Further, the tweets have been ranked based on relevancy. They used a threshold of 15

4 Baseline Results

4.1 Greta Thunberg

Baseline Used	Precision
Binary TF	*0.024*
Raw Count	*0.03*
Term Frequency TF	*0.03*
Log Normalization TF	*0.026*
Double Normalization TF	*0.028*

Precision @20

4.2 Virat Kohli

Baseline Used	Precision
Binary TF	*0.02*
Raw Count	*0.03*
Term Frequency TF	*0.03*
Log Normalization TF	*0.03*
Double Normalization TF	*0.0325*

Precision @20

5 References

https://scholarworks.uark.edu/cgi/viewcontent.cgi?article=1795&context=etd

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1aff7b429f99f529228a4299a5794971adeb1ca3

https://ieeexplore.ieee.org/document/7434278