News Recommendation System

Vishnumolakala Nikhila
MT21103
IIIT Delhi, India
vishnumolakala21103@iiitd.ac.in

Prachi Gupta
MT21135
IIIT Delhi, India
prachi21135@iiitd.ac.in

Nellore Bhavana
MT21131
IIIT Delhi, India
nellore21131@iiitd.ac.in

Madadi Vineeth Reddy
MT21046
IIIT Delhi, India
madadi21046@iiitd.ac.in

Munagala Krishna Sai
MT21054
IIIT Delhi, India
munagala21054@iiitd.ac.in

Elapanti Sri Sai Chaitanya
MT21120
IIIT Delhi, India
elapanti21120@iiitd.ac.in

1 ABSTRACT

Online consumption of news articles has gained popularity as the web feeds the reader with millions of news articles worldwide. The presented work proposes a personalized news recommendation that scrapes news from different sources and gives users recommendations based on their preferences.

Using cosine similarity measures, recommended feed is provided in the user-friendly interface where a model based on Support Vector Machine using TF-IDF vectorizer is proposed that categorizes the news articles based on distinct genres of news. So, along with the recommendation, different news categories are also displayed.

In an attempt to build a recommendation system with high accuracy, the presented work proposes a hybrid model based on content and collaborative filtering methods. The framework consists of 4 main components: news extraction, news classification, news recommendation, and a user-friendly portal to display the recommendations

2 INTRODUCTION

According to the Times Of India recent survey[1] about 68% of Indian users are consuming news articles from their smartphones, which is only going to increase with time. Today's world is flooded with information from all over the world. This phenomenon results in "Information overload." Information overload makes a person feel overwhelmed and it is hard to choose what to consume and what not to. Due to this, the reader wastes a lot of time and is often presented with news that might not interest him or doesn't suit his needs. We try to ease this problem by providing a personalized news recommendation system for the user based on his preferences, thus saving time and giving him the news he needs.

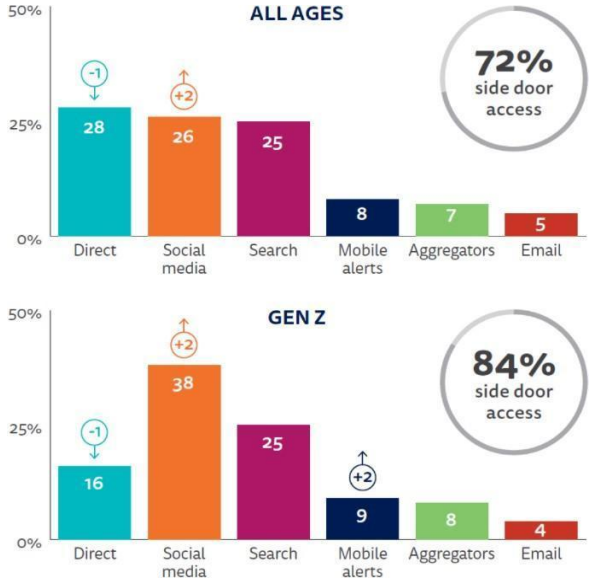

Figure 1: Survey report on the sources used by readers to consume their news (Source:digitalnewsreport.org)

As you can interpret from the above image, the number of readers who are reading directly from newspapers is decreasing gradually and are shifting towards online sources such as social media, web search, etc. Around 80% of the readers consume their news from non-physical newspapers.

Taking the reader experience further, many readers don't understand English and prefer reading news in their local language. So, we also provide news in multiple languages so that this system can reach the masses in their regional languages. Often, readers come across images and would like to know if there is any news regarding them. We also try to address this problem by providing the option of uploading the image and getting relevant information based on that image. We observed that users do not easily interpret many news websites as they are complex. We provide a simple web interface that any average person could use easily.

We not only provide personalized news but also provide news on hot topics and various categories such as politics, sports, etc. Thus making this a complete package for the reader, including both options of personalized news and news according to chosen categories.

3 PROBLEM STATEMENT

News Recommendation system helps gather news from various resources such as CNN-news, India Today, and The Hindu by web scraping. We present this news to the users based on their interests and multiple categories.

Two of the most popular methods for recommendation systems are content-based and collaborative-based. We recommend news to users based on their history in content-based filtering whereas collaborative-based filtering considers content read by other users having similar interests and recommends news based on those readers. We build our recommendation system that takes good care of both worlds by using a hybrid model that encompasses content and collaborative filtering to recommend news to the reader.

To view all the news gathered, we provide a simple web interface that contains various categories such as “Your recommendations” that present personalized news, “Hot news” that has the latest breaking news from all over the world, and various news according to categories. The news gathered is segregated into sports, education, business, entertainment, crime, politics, etc.

4 LITERATURE SURVEY

4.1 Multi-category news classification using Support Vector Machine based classifiers:

Saigal et al. [9] proposed a multi-category text classification based on Support vector machines and their different variants, specifically Least-squares Support Vector Machines, Twin Support Vector Machines, and Least-squares Twin Support Vector Machines (LS-TWSVM) classifiers. The performance of these classification algorithms is compared to benchmark datasets like Reuters and 20 Newsgroups. All these classifiers are binary classifiers and thus use the One-Against-All approach. LS-TWSVM outperformed the rest of the classifiers and achieved an accuracy of 92.96% since SVM classifiers outperformed the rest of the state-of-the-art models. Hence, the proposed methodology implemented SVM for news classification.

4.2 Content-based news recommendation:

Since there is a corpus of data that is available on the internet it is very difficult to filter and get the data in which the user is interested. To overcome this problem, in this paper they were using the cosine similarities for getting the personalized news for the user. Here, it finds the vector distances of the content that the user already read and the latest news available if it matches then that data is shown on the user feed.
This method was tested in an SME-FIIT research project with over ten thousand articles for which similarity was calculated in 2-3 seconds. [3]

4.3 News recommendation system using collaborative filtering method:

We have huge data for the news but in the unstructured format, So it's not possible to read everything which is available on the internet, it would be better if we get the data or the news in which we are interested in, and up-to-date information is presented online that can be accessed quickly, anytime, and anywhere. However, sometimes the news which is displayed is not up-to-date or not in accordance with the interests of the reader. So here in the paper [2] this problem we are overcoming by using the Collaborative Filtering (CF) to rank news as the best recommendation for readers. Here, using collaborative filtering, we get the data into our feed from a similar type of user. In this approach, we use Cosine-based Similarity to calculate the similarity of two users.

This method was tested on nineteen news and got an accuracy of 64.2% and an error value of 35.8%.

4.4 Semantics-based recommendation

TF-IDF along with cosine similarity is one of the common techniques used for news recommendation. The problem with this is that it doesn’t consider the actual meaning of the words. This paper [8] proposes two methods to solve this problem namely synset frequency and semantic similarity. Synset frequency doesn’t consider words directly but considers WordNet synonym sets. The semantic similarity makes use of five similarity measures to compute the similarity for news recommendations. The results have shown that these two methods are better than TF-IDF on f1-score.
The experiments were conducted on this method with the help of nineteen participants. SF-IDF (46.8%) outperforms TF-IDF (32.0%) on the F1-measure.

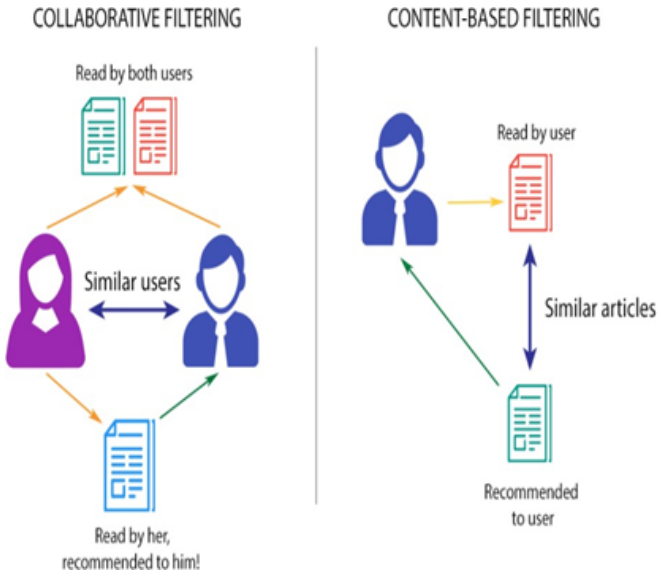

Figure 2: Collaborative filtering vs Content-based filtering

(source: google images)

4.5 MIND: A Large-scale Dataset for News Recommendation

In contrast to product and movie recommendations that are studied comprehensively, news recommendations don't have a high-quality dataset that is well labeled. So, this dataset [10] is made from user click logs of Microsoft news. This MIND dataset contains more than 1 million users and more than 160k news articles. It includes the title, abstract, and body. This dataset is considered to be a good test set for news recommendations.

4.6 Personalized News Recommendation and Simulation Based on Improved Collaborative Filtering Algorithm

Mostly, the news recommendation system follows collaborative-based filtering for the recommendation. The data preprocessing phase and the collaborative filtering at the nearest neighbor collection phase are enhanced to get more accurate recommendations. The enhancement combines collaborative-based filtering with the dichotomous K-means algorithm, and at the data preprocessing stage, a rating matrix is created that reduces the data sparsity. The effectiveness of the algorithm proposed in the paper [11] is tested by the simulation data set with recall rate and recommendation accuracy.Cosine similarity has given better results for the above algorithm than the Jaccard coefficient. So, further, we will be using cosine similarity.

4.7 HYPNER: A Hybrid Approach for Personalized News Recommendation

A similar framework is presented by Darvishy et al. [12] that aimed to resolve the scalability issue in the huge corpus of news sets available on the web, enhance the user's profile, display the features of diverse news items, and finally recommend the news articles. To compute the similarity ratio between a short-term user profile and news metadata, Cosine similarity and Jaccard similarity are used. A combined approach of content and collaborative filtering is used. With the help of HYPNER, news clustering, recommendations, and news selection become feasible. HYPNER is using a new clustering algorithm called Ordered clustering an unsupervised linear clustering algorithm. This paper used the F1 score as the metric and it achieved an 81.5% score.

5 BASELINE MODEL

5.1 Web Scraping

To deal with the tons of data present in this digital world, we use web scraping that collects the live and raw data from various news websites.

We are using the BeautifulSoup library to collect the raw data from the web. Here we collect the data from the IndiaToday website under the top stories category. We extract around 300 articles from India today on different days to get the recommendations aptly.

Organized the collected data into four columns article Id, Headline, description, and date. Then save the extracted data into a CSV file.

The data collected from these websites are preprocessed and used in the recommendation system.

5.2 Data Preprocessing

The data obtained from the web scraping is preprocessed by removing punctuation, stop words, stemming, and lemmatization. For the purpose of categorizing the news data, we collected the MIND dataset[13] and news-category dataset[14] from Kaggle and categorized them into 7 distinct categories (Sports, Education, Business, Entertainment, Crime, Politics, and Others). We used this customized dataset to build our model.

5.3 Models Used

5.3.1 Logistic Regression

Logistic regression is one of the statistical methods used to do binary classification like yes or no based on prior knowledge of the dataset. It predicts the outcome of a dependent variable based on the relationship between existing independent variables.

5.3.2 SVM Classifier

One of the most popular supervised learning algorithms which are used for classification and also regression problems is the support vector machine(SVM). The purpose of SVM is to create boundary lines to create n-dimensional space into classes so that we can get the class of the new point. The best boundary that we get is termed the hyperplane..

5.3.3 Random Forest

Random forest is a supervised algorithm that creates many decision trees and combines them into the forest with a single result. It uses ensemble learning, feature randomness, and bagging. Random forest is used for classification as well as regression.

5.3.4 Baseline results

MODEL	ACCURACY	PRECISION	RECALL
Logistic Regression	**079221**	**0.78920**	**079221**
SVM	**0.7987**	**0.7993**	**0.7987**
Random Forest	**0.74538**	**0.74257**	**0.74538**

5.4 Recommendation

To recommend the news articles to the user, we are taking the category wise classified scrapped news. We took the average word2vec vector of users and then compared this with all the collected news articles using cosine similarity. Finally, we are displaying the news articles in the descending order of the similarity. We are also recommending these articles category-wise.

6 PROPOSED SOLUTION

Personalized news recommendation system has 4 main components: News extraction, news classification, news recommendation, and a user-friendly portal to display the recommendations. In the first component, we scrap the latest news from multiple news sources (CNN-news, The Hindu, India Today). This data comes from a variety of sources so various text pre-processing techniques are applied like tokenization, lower-casing, stop word removal, lemmatization, and removing punctuation.

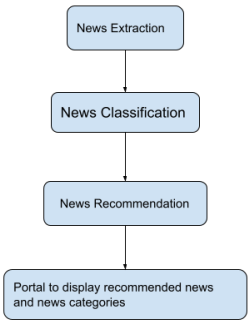

Figure 3: Components of the presented system

For news classification, we implemented a Support Vector Machine using a TF-IDF vectorizer that categorizes news into 7 distinct categories (Sports, Education, Business, Entertainment, Crime, Politics, and Others). The same categorized news will be displayed on the website under different feeds.

In the next component, the news is recommended using a hybrid approach. In content, the proposed approach is to take users' history by monitoring clicks and recommending further news based on the already read content but it leads to cold start problems. So, using a collaborative approach we aim to take new users' interests in the form of questionnaires. Also, it considers various readers of similar interests and recommends news based on those readers.

Problems with both the systems are eliciting user feedback. Hence, in this system, we are proposing a hybrid-based model that uses both content and a collaborative approach to give recommendations. Hybrid methods provide more accurate recommendations than pure approaches as they overcome the problems faced by these two approaches. Netflix is a popular example that uses a hybrid approach to recommend movies [7].

Using cosine similarity, the top 10 articles are recommended to the user. The same will be displayed in the "Recommendations" tab on the website along with other categories. Our system is not limited to only English-based articles but it also aims to provide multilingual news recommendations to the users to reach more audiences.

For the last component, we built a user-friendly website to display all the recommended news and news from different categories. For improvement, we also aim to collect user feedback after the users complete their visit to the website.

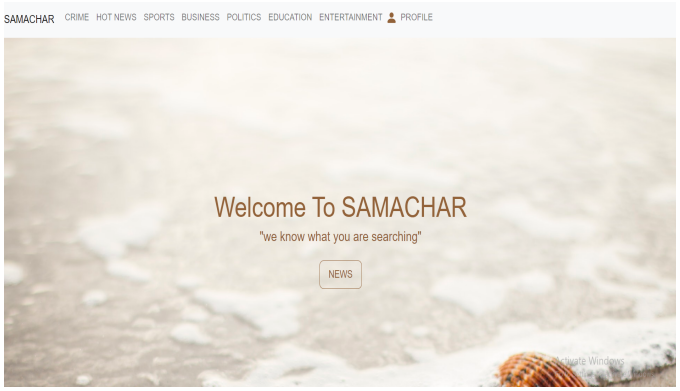

Figure 4: Home page of the web portal

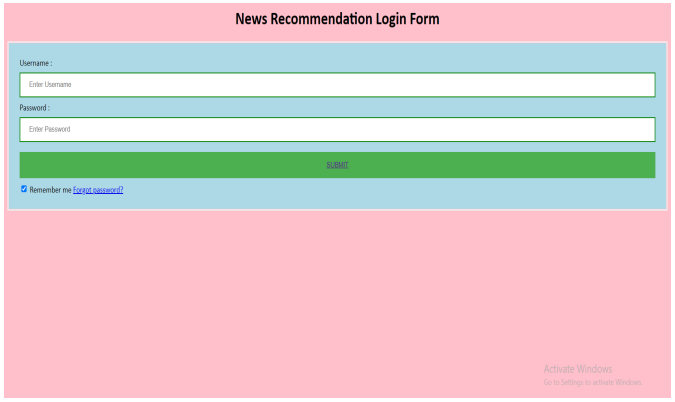

Figure 5: Login page

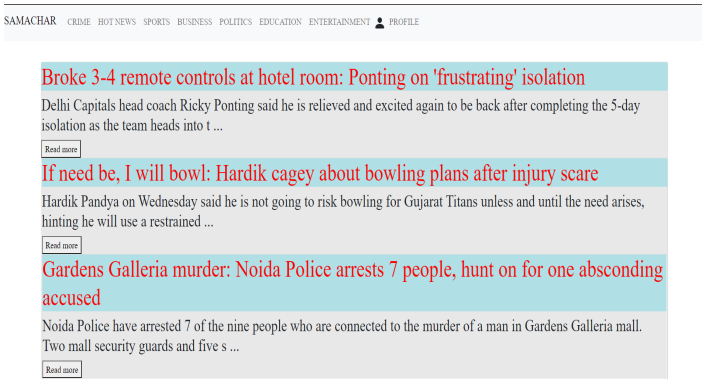

Figure 6: Recommended news articles

Recommendations: Top 10 recommendations to the user

	Unnamed: 0	Headline	Description	Cosine Simil
0	1	Punjab CM Bhagwant Mann slams opposition for c...	Punjab Chief Minister Bhagwant Mann on Tuesday...	0.9
1	2	Myanmar junta court convicts Aung San Suu Kyi ...	A court in military-ruled Myanmar's capital Na...	0.9
2	3	Patna to reel under heatwave conditions till F...	Bihar's capital Patna is likely to remain in t...	0.9
3	4	Blood of Chinese cannot be shed in vain, perpe...	China on Wednesday asked Pakistan to scale-up ...	0.9
4	5	Palm oil export ban by Indonesia begins tomorr...	Indonesia, the world's top palm oil producer, ...	0.9
5	6	Beast box office collection Day 14	Thalapathy Vijay's Beast has slowed down consi...	0.9
6	7	Bengaluru: Nearly three lakh citizens yet to t...	Even as the fear of a fourth Covid wave looms ...	0.9
7	8	Watch LIVE | SpaceX launches four astronauts t...	Four astronauts part of the Crew-4 mission wer...	0.9
8	9	Pune farmer hires helicopter to accord grand w...	A farmer from Maharashtra's Pune district hire...	0.9
9	10	A mother of 2 and wife of a doctor, Karachi su...	The woman who carried out the Karachi suicide ...	0.9

Figure 8: recommendation given to user

Novelty: Multi-lingual translation (in backend)

	Unnamed: 0	Headline	Description	Cosine S
0	1	पंजाब के सीएम भगवंत मान ने दिल्ली के साथ एमओयू...	पंजाब के मुख्यमंत्री भगवंत मान ने मंगलवार को व...	0.9999726
1	2	म्यांमार की जुंटा अदालत ने आंग सान सू की को भ्...	सैन्य शासित म्यांमार की राजधानी नेपीताव की एक ...	0.9999716
2	3	पटना शुक्रवार तक लू की चपेट में रहेगा, मौसम वि...	बिहार की राजधानी पटना के अगले तीन दिनों तक लू ...	0.99999712
3	4	चीनियों का खून व्यर्थ नहीं बहाया जा सकता, दोषि...	चीन ने बुधवार को पाकिस्तान से देश में काम करने...	0.9999701
4	5	इंडोनेशिया द्वारा पाम तेल निर्यात प्रतिबंध कल ...	इंडोनेशिया, दुनिया का शीर्ष पाम तेल उत्पादक, र...	0.9999694
5	6	बीस्ट बॉक्स ऑफिस कलेक्शन दिन 14	थलपति विजय का जानवर अपने दूसरे सप्ताह में काफी...	0.9999692
6	7	बेंगलुरु: लगभग तीन लाख नागरिकों ने अभी तक दूसर...	यहां तक कि कर्नाटक में चौथी कोविड लहर का ड...	0.9999692
7	8	लाइव देखें | स्पेसएक्स ने फाल्कन-9 . पर अंतरिक...	क्रू-4 मिशन के चार अंतरिक्ष यात्रियों को बुधवा...	0.9999691
8	9	पुणे के किसान ने नवजात पोती का भव्य स्वागत करन...	महाराष्ट्र के पुणे जिले के एक किसान ने अपनी नव...	0.9999690
9	10	दो बच्चों की मां और एक डॉक्टर की पत्नी, कराची ...	जिस महिला ने कराची आत्मघाती बम विस्फोट को अंजा...	0.9999680

Figure 9: Translated news recommendations in Hindi

	Unnamed: 0	Headline	Des
0	1	ధిల్లీతో కుదుర్చుకున్న ఎంఒయూపై విపక్షాలు విమర్...	ఆప్ ప్రభుత్వం ఇటీవలి కీలక నిర్ణయాలను
1	2	అవినీతి కేసులో ఆంగ్ సాన్ సూకీని దోషిగా నిర్ధార...	మిలటరీ పాలనలో ఉన్న మయన్మార్ రాజధాని నై
2	3	పాట్నా శుక్రవారం వరకు హీట్‌వేవ్ పరిస్థితులలో క...	బీహార్ రాజధాని పాట్నా రాబోయే మూడు రోజుల
3	4	చైనీయుల రక్తం వృథా కాదు, దోషులను శిక్షించాలి: ...	కరాచీ యూనివర్సిటీలో ముగ్గురు చైనీస్ అధ
4	5	ఇండోనేషియా ద్వారా పామాయిల్ ఎగుమతి నిషేధం రేపటి...	ప్రపంచంలోని అగ్రశ్రేణి పామాయిల్ ఉత్పత్తి
5	6	బీస్ట్ బాక్స్ ఆఫీస్ కలెక్షన్ డే 14	దళపతి విజయ్ మృగం రెండవ వారంలో బాగా నెమ
6	7	బెంగళూరు: దాదాపు మూడు లక్షల మంది పౌరులు ఇంకా ర...	నాల్గవ కోవిడ్ వేవ్ కర్ణాటకపై భయంతో కూడ
7	8	లైవ్ చూడండి | స్పేస్‌ఎక్స్ నలుగురు వ్యోమగాములన...	క్రూ-4 మిషన్‌లో భాగంగా నలుగురు వ్యోమగాము
8	9	అప్పుడే పుట్టిన మనవరాలికి ఘన స్వాగతం పలికేందుక...	మహారాష్ట్రలోని పూణె జిల్లాకు చెందిన ఒ రై
9	10	ఇద్దరు పిల్లల తల్లి మరియు డాక్టర్ భార్య, కరాచీ...	కరాచీ ఆత్మాహుతి బాంబు దాడికి పాల్పడిన మ

Figure 10: Translated news recommendations in Telugu

7 CHALLENGES

The proposed methodology is experimented with and evaluated on datasets created using the MIND dataset and the Kaggle news category dataset. For the user interface, real-time scrapping of data is carried out at regular intervals. So, recommendations take time to display once the user reloads the webpage. For the time being our focus lies on providing personalized recommendations to the user so, we didn't provide images along with the titles and descriptions of the news articles.

We provided translated news articles functionality in the backend where the input is the scrapped news and recommendations are given in the desired language. But this functionality is not displayed in the frontend currently. Another limitation is maintaining the privacy of the user as the user's history is also taken into account.

8 FUTURE SCOPE

More advanced methods can be implemented to combine the two popular approaches i.e., content and collaborative filtering to leverage both the approaches. Along with news websites, we can consider different news pieces from social media as users spend a considerate amount of time on social media platforms. More optimized approaches can be implemented to reduce the computation time for recommending news sets to the user in real-time.

Constructing effective models to identify influencers and formulating the frameworks based on that is also a vision as influencers have a significant impact on information diffusion.

9 CONCLUSION

This approach automates the process of providing the most relevant information from authorized news websites. The project focuses mainly on providing news recommendations based on users' needs and thus reducing the manual effort to seek relevant information. Previous recommender systems fetched the data from already stored data sets, not necessarily live news.Thus we propose an application that recommends the right news at the right time to the user.

This project proposed an SVM-based model implemented using TF-IDF to categorize the news articles and later recommends the top 5 news using cosine similarity. SVM is considered among others baseline models as it achieved better performance as compared to other datasets.

To achieve high accuracy and a more personalized feed to the user, a hybrid approach utilizing both content and a collaborative filtering approach are used.

10 REFERENCES

[1] T. C. / U. Mar 25, 2019, and 19:08 Ist, "68% of Indian users consume news on smartphones: Report | India News - Times of India," *The Times of India*. https://timesofindia.indiatimes.com/india/68-of-indian-users-consume-news-on-smartphones-report/articleshow/68565146.cms

[2] A Wahana et al 2019 J. Phys.: Conf. Ser. 1402 077010 https://iopscience.iop.org/article/10.1088/1742-6596/1402/7/077010/pdf

[3] Kompan, M., Bieliková, M. (2010). Content-Based News Recommendation. In: Buccafurri, F., Semeraro, G. (eds) E-Commerce and Web Technologies. EC-Web 2010. Lecture Notes in Business Information Processing, vol 61. Springer, Berlin,Heidelberg.

[4] Ismail, Walaa & Nasr, Mona & Saied, Mohamed. (2018). A HYBRID NEWS RECOMMENDER SYSTEM.

[5] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In Proceedings of the 15th international conference on Intelligent user Interface (IUI '10). Association for Computing Machinery, New York, NY, USA, 31–40.

[6] Z. Xia, S. Xu, N. Liu, and Z. Zhao, "Hot News Recommendation System from Heterogeneous Websites Based on Bayesian Model," The Scientific World Journal, vol. 2014, p. e734351, Jun. 2014.

[7] A. Yaman, "Hybrid Recommender System-Netflix Prize Dataset," *CodeX*, Mar. 30, 2021. https://medium.com/codex/hybrid-recommender-system-netflix-prize-dataset-e9f6b4a875aa.

[8] Michel Capelle, Flavius Frasincar, Marnix Moerland, and Frederik Hogenboom. 2012. Semantics-based news recommendation. In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS '12). Association for Computing Machinery, New York, NY, USA, Article 27, 1–9.

[9] Saigal, P., Khanna, V. Multi-category news classification using Support Vector Machine based classifiers. *SN Appl. Sci.* 2, 458 (2020).

[10] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3597–3606, Online. Association for Computational Linguistics.

[11] Kunni Han, "Personalized News Recommendation and Simulation Based on Improved Collaborative Filtering Algorithm", Complexity, vol. 2020, Article ID 8834908, 12 pages, 2020.

[12] A. Darvishy, H. Ibrahim, F. Sidi and A. Mustapha, "HYPNER: A Hybrid Approach for Personalized News Recommendation," in IEEE Access, vol. 8, pp. 46877-46894, 2020. .

[13] “MIND: Microsoft News Recommendation Dataset,” www.kaggle.com. https://www.kaggle.com/datasets/arashnic/mind-news-dataset.

[14] “News Category Dataset,” *www.kaggle.com*. https://www.kaggle.com/datasets/rmisra/news-category-dataset.