IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.005, LIIR302
DOI: https://doi.org/10.52731/liir.v005.302

Investigation of the Diversity of News Reading Articles and Browsing Trends Using Sentence-BERT

Akihisa Takiguchi *, Tsunenori Mine *, Yutaka Arakawa *

Abstract

In modern life, platforms like a social networking service (SNS) play a crucial role by utilizing recommender systems to present useful information from vast datasets. However, the advancement of these systems has led to biases in user-exposed information, causing societal issues like public opinion conflicts and defamation. Furthermore, sentiment biases in the information viewed contribute to this problem, described as "informational malnutrition". This highlights the need for "informational health", where user access to various information maintains the balance of information intake they seek.

In this work, we explored differences in user viewing tendencies based on the diversity of viewed articles, employing a dataset of news articles and user logs. We utilized Sentence-BERT, a natural language processing model known for its effective sentence similarity analysis, to vectorize articles and score their similarity, measuring users' article diversity. Our analysis, considering sentiment content biases, used multiple regression. The results suggest that users with diverse viewing habits tend to prefer articles with a negative bias and that news in categories such as music, current affairs, and politics have a low contribution to the diversity of information viewed, and conversely, categories like entertainment and lifestyle content tend to have a high contribution to the diversity of information viewed.

Keywords: Diversity, News, Recommender System, Sentence Embedding

1 Introduction

In modern life, platform services such as a social networking service (SNS) have become indispensable. These services utilize recommender systems to select and present information that is deemed useful to users from a large pool of data. As a result, it has become easier for users to access the information they prefer. However, the development of these recommender systems has led to phenomena such as echo chambers, where users are only exposed to opinions similar to their own, and filter bubbles, where information that does not match the user's preferences is automatically excluded, making it easier for biases in the viewed information to arise [1]. These issues contribute to problems such as cyber cascades, which are a form of group polarization leading to societal issues like public opinion division and conflict, defamation, and even criminal cases [2]. Furthermore, studies on the

* ISEE, Kyushu University, Fukuoka, Japan

impact of negative news articles have reported that while negative news articles can have adverse mental effects, they also accelerate the consumption of negative news [3], enhancing the tendency towards filter bubbles [4], raising similar concerns. Toriumi et al. [5] have described this state of biased information consumption as "informational malnutrition", comparing it to dietary intake, and have issued a warning about the current systems of information presentation and user behavior. Moreover, to increase long-term usage of services, the construction of recommender systems that incorporate a diversity of information is deemed necessary, and research is being conducted toward this end [6],[7]. Thus, in recent years, the diversity of information presented to users has been considered important from the perspectives of societal issues, sentiment impact of information, and user engagement. It is posited that exposing users to a diverse and broad range of information can lead to a state of "informational health", where the intake of information is balanced [8]. To achieve this, it is necessary to understand users' tendencies in consuming information from the viewpoint of information diversity.

This study focuses on news sites within platform services and investigates the differences in viewing tendencies according to the diversity of viewed articles using a large-scale dataset comprising news articles and user browsing logs. Previous research has vectorized news articles to measure the diversity of user-viewed articles. These prior methods, based on utilizing labels of clusters assigned to users based on past click information [9], have not employed natural language processing models to measure diversity. Moreover, the dataset of news articles used in [10] suggests that the method of extracting features using the natural language processing model BERT [11] is effective. Sentence-BERT, which is fine-tuned based on BERT to achieve high precision in measuring semantic similarity between sentences, aims to map similar sentences to closer vector spaces and dissimilar ones further apart. It has been demonstrated through research that Sentence-BERT outperforms traditional BERT models, especially in tasks such as sentence-level similarity estimation and semantic search [12]. These characteristics are considered useful for measuring and analyzing the similarity between news articles, allowing for a more detailed understanding of semantic differences between articles.

Therefore, this study utilized the Sentence-BERT model to vectorize news articles and score the similarity to measure the diversity score of users' viewed articles. Based on the diversity score of viewed articles and considering the adverse effects of biased sentiments in news articles, the analysis of users' viewing tendencies was conducted from the following two questions:

- RQ1: What is the relationship between the diversity and sentiment of articles viewed?
- RQ2: Depending on the degree of diversity of the articles viewed, what differences exist in the trends of the categories viewed?

For these investigations, sentiment analysis of news articles was performed first, followed by scoring the sentiment content of the articles consumed by users. A multiple regression analysis using the least squares method was conducted on the sentiment score and the diversity score of viewed articles, suggesting that the more negative the viewed articles, the higher the diversity. Similarly, an analysis of the relationship between the diversity score of the articles viewed and the trends of the categories indicated that users with higher diversity in their viewing habits tend to view news related to music, current events, politics, crime, world situations, science and technology, finance, and are less likely to view entertainment, food and drink, movies, and weather news. Additionally, both analyses confirmed

that users who view more articles tend to have a higher diversity in their viewed articles, consistent with previous research [13].

2 Related Work

2.1 The Negative Impacts of Biased Information and Sentiments

In the services of the digital platform today, an enormous exchange of information occurs daily. Recommender systems select and present information deemed useful to users based on their objectives, facilitating comfortable access to preferred information. However, from a business perspective, it has been found that recommender systems can overly cater to user preferences, continuously recommending similar content and exposing users to biased information [13]. Toriumi et al. liken this condition to "informational malnutrition," equating the consumption of information with food intake, and raising an alarm about modern information presentation systems and user behavior [5].

In addition, research has been conducted that focuses on the sentiment aspect of biased information, particularly the impact of negative news articles. Ytre et al.'s study discusses the relationship between mental health and news usage during the COVID-19 pandemic, reporting that negative news articles can have detrimental mental effects while accelerating the consumption of negative news [3]. Moreover, Robertson et al. reported that negative news articles increase user viewership rates, especially for political and economic articles, suggesting that increased consumption of negative news articles inadvertently exposes users to content that can cause political divisions and polarization [14]. The study by Ohata et al. found that users who click on negative news articles are more likely to be recommended negative articles, making them more susceptible to consumption compared to other articles, thus intensifying the filter bubble effect [4]. Therefore, the presentation of biased information can have significant social impacts, necessitating consideration of the sentiment aspect of the information in its analysis.

2.2 Diversity of Viewed Information

Various studies have been conducted on the diversity of viewed information. Anderson et al.'s research proposed the Generalist-Specialist Score (GS-Score) as a measure of user activity diversity, using this score to analyze user communities on online platforms [15]. Applying the GS-Score, Anderson et al. also analyzed the diversity of user preferences and recommender systems on the online music streaming service Spotify, reporting that users with a higher diversity of listened tracks exhibited higher user engagement and continuous usage [7]. Similarly, Holtz et al. [16] also conducted research on the diversity of listened to tracks and user engagement on Spotify. Additionally, Abdool et al. conducted research incorporating diversity into the search algorithm of the lodging rental web service Airbnb, suggesting that displaying diverse search results can widen users' choices and enhance user experience [6]. Thus, increasing the diversity of viewed information is considered important for long-term user engagement with services.

2.3 Diversity of News Viewing Articles

Research on the diversity of information viewed in the field of news viewing analysis. Sasaki et al. analyzed the diversity of articles viewed and user behavior using the usage

	id	category	subcategory	title	abstract	url	title_entities	abstract_entities
0	N55528	lifestyle	lifestyleroyals	The Brands Queen Elizabeth, Prince Charles, an...	Shop the notebooks, jackets, and more that the...	https://assets.msn.com/labs/mind/AAGH0ET.html	[{"Label": "Prince Philip, Duke of Edinburgh",...	[]
1	N18955	health	medical	Dispose of unwanted prescription drugs during ...	NaN	https://assets.msn.com/labs/mind/AAISxPN.html	[{"Label": "Drug Enforcement Administration", ...	[]
2	N61837	news	newsworld	The Cost of Trump's Aid Freeze in the Trenches...	Lt. Ivan Molchanets peeked over a parapet of s...	https://assets.msn.com/labs/mind/AAJgNsz.html	[]	[{"Label": "Ukraine", "Type": "G", "Wikidatald...
3	N53526	health	voices	I Was An NBA Wife. Here's How It Affected My M...	I felt like I was a fraud, and being an NBA wi...	https://assets.msn.com/labs/mind/AACk2N6.html	[]	[{"Label": "National Basketball Association", ...
4	N38324	health	medical	How to Get Rid of Skin Tags, According to a De...	They seem harmless, but there's a very good re...	https://assets.msn.com/labs/mind/AAAKEkt.html	[{"Label": "Skin tag", "Type": "C", "Wikidatal...	[{"Label": "Skin tag", "Type": "C", "Wikidatal...
...	...	...	...	...	...	...	...	...

Figure 1: List of News Article Data (Excerpts)

logs of “Yahoo! JAPAN”, a portal site operated by LY Corporation, specifically its news site “Yahoo News”, employing the GS-Score as an indicator of the diversity of user articles [12]. Similarly, Suganuma et al. conducted analysis using the GS-Score on user behavior logs from the news application “Gunosy” provided by Gunosy Inc. [8]. Both studies analyzed the relationship between the diversity of news articles and user engagement, reporting that users consuming a wide range of information have higher continuation rates, i.e., higher user engagement. However, no analyses have been conducted on the relationship between the diversity of users’ news viewing articles and the specific content trends of those articles. Additionally, while traditional research has vectorized news articles to measure the diversity of users’ viewed information, no study has been found that uses natural language processing models to measure the diversity score of news viewing articles.

In light of this, the current study employs Sentence-BERT, a natural language processing model that demonstrates superior performance in textual similarity analysis compared to the traditional BERT model, for the embedding of news articles. This approach is based on the effectiveness of using BERT to extract characteristics in the analysis of datasets from news articles [9]. Following considerations in Section 2.1, sentiment analysis of news articles was performed to investigate the relationship between the diversity of news viewing articles and sentiments. Subsequently, the study examined the differences in specific browsing category trends according to the degree of diversity in users’ viewed articles.

3 Diversity of Viewed Articles

3.1 Dataset

This study uses MIND (Microsoft News Dataset), an open dataset that was collected over six weeks from October 12 to November 22, 2019, consisting of anonymized behavior logs provided by Microsoft for Microsoft News. MIND contains approximately 160,000 English news articles and over 15 million impression logs generated by about 1 million users. Each news article includes the title, summary, body, category, entities, etc., as shown in Figure 1. Each user’s impression log contains the results of news article recommendations and the user’s previous browsing history, as shown in Figure 2. For this research, the MIND-small dataset, a relatively small dataset, was used, containing 42,416 news articles and the browsing history of 33,948 users who have viewed at least one article.

	user_id	time	history	impressions
0	U80234	11/15/2019 12:37:50 PM	N55189 N46039 N51741 N53234 N11276 N264 N40716...	N28682-0 N48740-0 N31958-1 N34130-0 N6916-0 N5...
1	U60458	11/15/2019 7:11:50 AM	N58715 N32109 N51180 N33438 N54827 N28488 N611...	N20036-0 N23513-1 N32536-0 N46976-0 N35216-0 N...
2	U44190	11/15/2019 9:55:12 AM	N56253 N1150 N55189 N16233 N61704 N51706 N5303...	N36779-0 N62365-0 N58098-0 N5472-0 N13408-0 N5...
3	U87380	11/15/2019 3:12:46 PM	N63554 N49153 N28678 N23232 N43369 N58518 N444...	N6950-0 N60215-0 N6074-0 N11930-0 N6916-0 N248...
4	U9444	11/15/2019 8:25:46 AM	N51692 N18285 N26015 N22679 N55556	N5940-1 N23513-0 N49285-0 N23355-0 N19990-0 N3...
...	...	...	...	...

Figure 2: List of User Log Data (Excerpts)

3.2 News Article Embedding Method

The study employs Sentence-BERT, proposed by Nils et al., for embedding news articles. Sentence-BERT is a variant of the natural language processing model BERT, enhanced with a pooling layer on top of the pre-trained BERT model, utilizing a "Siamese Network" approach for more accurate embedding. This model fine-tunes using a loss function called Contrastive Loss to map similar sentences closer in vector space and dissimilar sentences further apart. Especially, Sentence-BERT has been shown to perform over 20% better on average than the traditional BERT model in benchmarks such as STS for sentence-level similarity estimation and semantic search tasks. This feature is useful in analyzing the similarity between news articles, allowing for a more detailed understanding of semantic differences between articles.

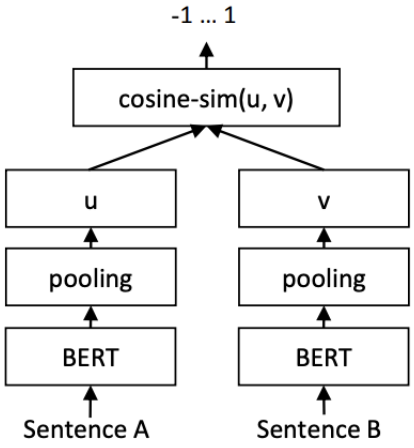

Figure 3: Sentence-BERT Architecture at Inference

Among the models of Sentence-BERT, paraphrase-MiniLM-L6-v2[1] is used for embedding the combined text of news article titles and summaries. This model, a fine-tuned version of Microsoft's pre-trained model MiniLM-L12-H384-uncased[2], outputs a 384-dimensional vector. The embedding results with this model are visualized by reducing the dimensions of news article vectors with t-SNE, as shown in Figure 4.

3.3 Definition of the Diversity of Viewed Articles (GS-Score)

Using the embedding described in the previous section, GS-Score is introduced as an indicator to represent the diversity of articles viewed by a user. The GS-Score of a user u, who has read a list of articles $\{s_1, ..., s_i, ...\}$ during a period, is defined as follows:

$$GS(u) = \frac{1}{\sum_i w_i} \sum_i w_i \frac{\vec{s_i} \cdot \vec{\mu}}{|\vec{s_i}||\vec{\mu}|} \tag{1}$$

[1] `https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2`
[2] `https://huggingface.co/microsoft/MiniLM-L12-H384-uncased`

where $\vec{s_i}$ is the vector representation of the article s_i, w_i is the number of times the article s_i was read, and $\vec{\mu}$ is the centroid vector of the user's viewed articles, determined by the equation:

$$\vec{\mu} = \frac{1}{\sum_i w_i} \sum_i w_i \vec{s_i} \tag{2}$$

In this study's dataset, each article is recorded only once per user, so $w_i = 1$ for all i, simplifying the GS-Score and $\vec{\mu}$ as follows:

$$GS(u) = \frac{1}{N} \sum_i \frac{\vec{s_i} \cdot \vec{\mu}}{|\vec{s_i}||\vec{\mu}|} \tag{3}$$

$$\vec{\mu} = \frac{1}{N} \vec{s_i} \tag{4}$$

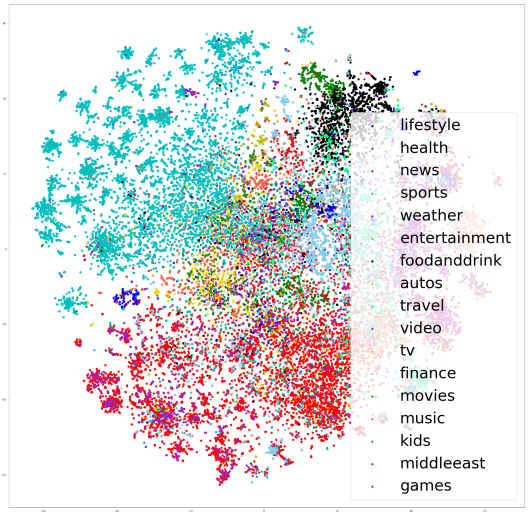

Figure 4: Result of Dimensionality Reduction of a Vector of News Articles by t-SNE

GS-Score ranges from -1 to 1, where values closer to 1 indicate homogeneity, and values closer to -1 indicate diversity in viewed articles.

3.4 Relationship between GS-Score and Number of Viewed Articles

According to Anderson et al., GS-Score is influenced by the number of items considered. Sasaki et al. investigated the impact of the number of news viewing articles on GS-Score, reporting a decrease in GS-Score with an increase in the number of articles viewed, asymptotically approaching a certain value. This study similarly conducted a preliminary experiment to confirm the impact of the number of articles viewed by users on GS-Score, visualizing the distribution of users' GS-Scores in Figure 5 and the relationship between GS-Score and the number of viewed articles in Figure 6. The results confirmed a similar phenomenon, where the GS-Score asymptotically approaches around 0.85 as the number of viewed articles increases, suggesting a bias in the embedding of news articles used in the dataset.

Considering these preliminary experiments, the analysis includes the user's number of viewed articles as an explanatory variable, along with sentiment scores and browsing categories, to analyze how browsing tendencies affect the diversity of viewed articles represented by GS-Score through multiple regression analysis. The analysis specifically focused on users who viewed more than 10 articles to account for potential fluctuations in GS-Score due to limited or incidental site use. The mean, median, and standard deviation of GS-Scores for all users and those with more than 10 viewed articles are shown in Table 1, with the distribution illustrated in Figure 3.7. Compared to the overall user base, users with more than 10 articles viewed show lower average, median, and standard deviation values for GS-Score, indicating a narrower distribution and lower diversity.

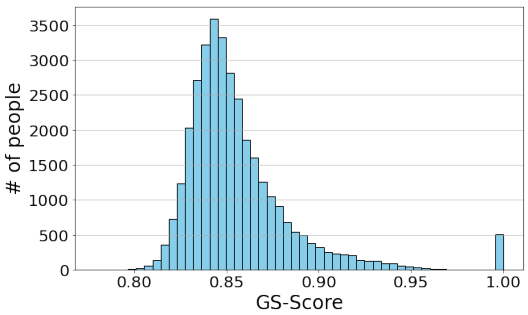

Figure 5: Distribution of Users' GS-Scores

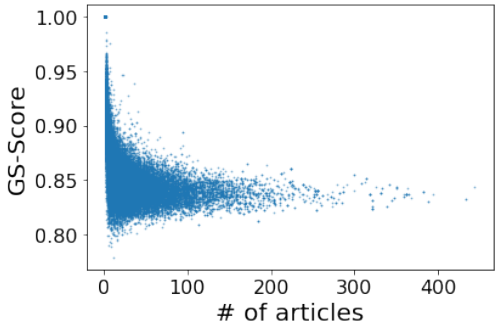

Figure 6: Distribution of Users' GS-Scores and Number of Articles Viewed

Table 1: Mean, Median, and Standard Deviation of User's GS-Scores

	Mean	Median	Standard Deviation
Overall	0.856	0.849	0.0301
More than 10 articles viewed	0.845	0.844	0.0148

4 The Relationship Between Diversity and Sentiment of Viewed Articles (RQ1)

4.1 Sentiment Scoring of Viewed Articles

4.1.1 Sentiment Analysis Model

For the sentiment analysis of news articles, the model lxyuan/distilbert-base-multilingual-cased-sentiments-student[3], based on DistilBERT—a lightweight version of BERT—was used. This model fine-tunes DistilBERT for sentiment analysis through zero-shot classification on a multilingual sentiment dataset, producing scores for positive, neutral, and negative sentiments ranging from 0 to 1 for a given text input as a probability distribution.

4.1.2 User Sentiment Score

Using the model described in Section 4.1.1, and similarly to Section 3.2, the titles and summaries of the news articles were inputted to measure the sentiment scores for each news article. The sentiment score for news articles was calculated as the difference between the positive and negative scores. A mean sentiment score of viewed articles was computed for each user, termed as the user's sentiment score. The mean, median, and standard deviation of the sentiment scores for news articles and users are shown in Table 2, with the distribution illustrated in Figure 8.

Both news articles and users show a tendency towards negative values in the mean and median sentiment scores, suggesting a preference for negative news articles. This corroborates the findings of Robertson et al. [14].

[3]`https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student`

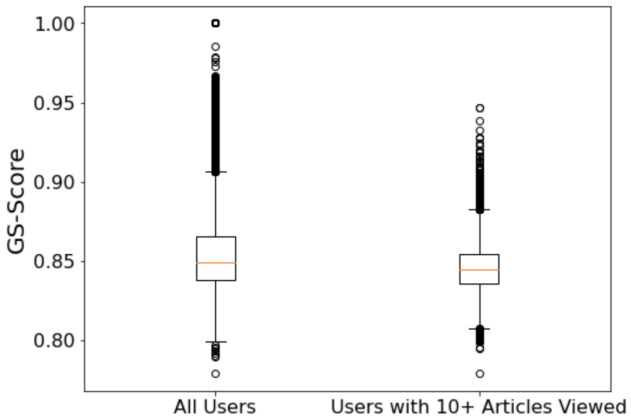

Figure 7: GS-Score Distribution for All Users and Users Who Viewed More than 10 Articles

Table 2: Mean, Median, and Standard Deviation of Sentiment Scores

	Mean	Median	Standard Deviation
News Articles	-0.074	-0.100	0.434
Users	-0.099	-0.103	0.140

4.2 Results and Discussion

A multiple regression analysis was conducted for 24,404 users with more than 10 viewed articles, using the GS-Score as the dependent variable and the user's sentiment score and the number of viewed articles as independent variables. The results are shown in Table 3 and Figure 9 ($R^2 = 0.152, F = 2186.10, p < 0.01$).

As shown in Table 3, both the user's sentiment score and the number of viewed articles were significantly related to the GS-Score. A higher sentiment score was associated with a higher GS-Score ($p < 0.01$), and a greater number of viewed articles was associated with a lower GS-Score ($p < 0.01$). The trend of a lower GS-Score with a higher number of viewed articles aligns with the results of the preliminary experiment in Section 3.4. This suggests that the more positive the user's viewing tendencies, the lower the diversity of viewed articles, or conversely, the more negative the user's viewing tendencies, the higher the diversity of viewed articles.

5 The Relationship Between Diversity of Viewed Articles and Browsing Category Trends (RQ2)

5.1 News Categories in the Dataset

The breakdown of news article categories within the dataset used for this analysis is as shown in Table 4. There are a total of 17 news article categories, with 42,416 articles recorded. The categories 'news' and 'sports' together make up half of the dataset. Additionally, the top 10 subcategories within the 'news' category are also listed in Table 5. The 'news' category mainly consists of articles related to US domestic affairs (newsus), politics (newspolitics), crime (newscrime), world affairs (newsworld), and science and technology

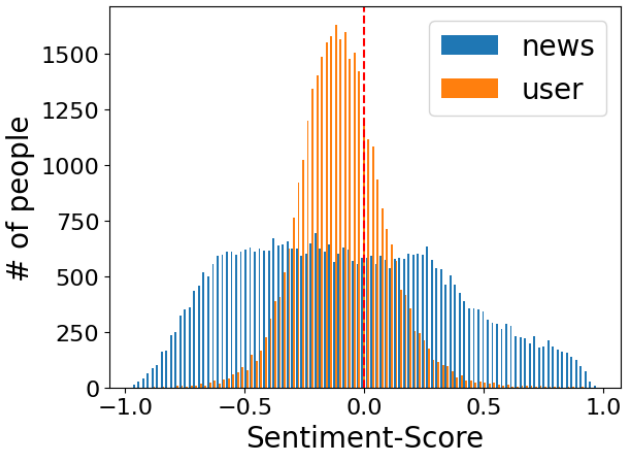

Figure 8: Distribution of Sentiment Scores

Table 3: Results of the Multiple Regression Analysis Model with Sentiment Score and Number of Viewed Articles as Explanatory Variables

Data Item	Coefficient	95% Confidence Interval	t-value	p-value
Sentiment Score	**0.0301**	[0.0289, 0.0313]	50.32	< 0.01
Number of Viewed Articles	**-0.000087**	[-0.000091, -0.000083]	-41.30	< 0.01
(Intercept)	0.8522	[0.8519, 0.8525]	6119.18	< 0.01

(newsscienceandtechnology).

5.2 Analysis Method

To analyze users' browsing category trends, the number of articles each user read in each category was normalized by dividing by the total number of articles read by that user. Based on this, a multiple regression analysis was conducted with the proportion of articles read in each category and the total number of articles read as explanatory variables, and GS-Score as the dependent variable. Due to the large number of browsing category items, multiple regression analysis using all categories as explanatory variables could lead to multicollinearity. Therefore, a stepwise method based on a p-value threshold of 0.05 was used to select significant categories as explanatory variables. As a result, categories 'kids', 'middleeast', and 'games' were deemed not significant and were excluded from the analysis.

5.3 Results and Discussion

The results of the multiple regression analysis conducted with GS-Score as the dependent variable and the proportions of category articles read, selected through stepwise method, along with the total number of articles read as explanatory variables, are shown in Table 6, and Figure 10 ($R^2 = 0.197, F = 93.37, p < 0.001$). The table reveals that both the proportion of articles read in each category and the total number of articles read have a significant relationship with GS-Score. Categories such as 'news', 'sports', 'finance', 'health', 'music' showed smaller effects on GS-Score, while 'foodanddrink', 'video', 'weather', 'entertainment' showed larger effects ($p < 0.05$). Additionally, a trend similar to the findings in the

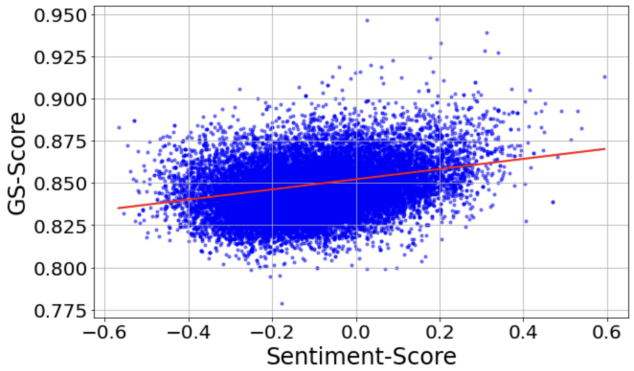

Figure 9: Scatterplot of GS-Scores and Sentiment Scores

Table 4: Breakdown of Categories

Category	Number of Articles
news	13,043
sports	11,760
finance	2,563
foodanddrink	2,248
lifestyle	2,132
travel	1,845
health	1,715
video	1,600
autos	1,490
weather	1,462
tv	822
music	625
entertainment	559
movies	538
kids	11
middleeast	2
games	1
Total	42,416

Table 5: Top 10 subcategories of 'news'

Subcategory	Number of Articles
newsus	5,335
newspolitics	2,398
newscrime	1,797
newsworld	1,525
newsscienceandtechnology	971
newsopinion	305
newsoffbeat	285
newsgoodnews	179
elections-2020-us	66
newsbusiness	49
Total	12,910

supplementary experiment of Section 3.4 was confirmed, where an increase in the number of articles read leads to a decrease in GS-Score ($p < 0.01$).

From these findings, it is suggested that news in categories such as music, current affairs, politics, crime incidents, world affairs, science and technology, and finance have a low contribution to the diversity of information viewed. Conversely, categories like entertainment, food and dining, movies, and weather tend to have a high contribution to the diversity of information viewed.

6 Discussion

In this study, we utilized the GS-Score as an indicator to represent the diversity of articles viewed by users in our analysis of user browsing tendencies. While previous research has shown GS-Scores ranging from approximately 0.3 to 1.0 [7], in our study, the scores were relatively narrowly distributed between approximately 0.8 to 1.0, indicating that the diver-

Table 6: Results of the Multiple Regression Analysis Model with the Proportion of Browsing Categories as Explanatory Variables

Data Item	Coefficient	95% Confidence Interval	t-value	p-value
news	**0.3589**	[0.037, 0.681]	2.187	0.029
sports	**0.3640**	[0.042, 0.686]	2.218	0.027
finance	**0.3593**	[0.038, 0.681]	2.190	0.029
foodanddrink	**0.3891**	[0.067, 0.711]	2.371	0.018
lifestyle	0.3809	[0.059, 0.703]	2.321	0.020
travel	0.3840	[0.062, 0.706]	2.340	0.019
health	**0.3625**	[0.041, 0.684]	2.209	0.027
video	**0.3894**	[0.068, 0.711]	2.372	0.018
autos	0.3779	[0.056, 0.700]	2.303	0.021
weather	**0.3892**	[0.067, 0.711]	2.371	0.018
tv	0.3701	[0.048, 0.692]	2.256	0.024
music	**0.3486**	[0.027, 0.670]	2.124	0.034
entertainment	**0.4256**	[0.104, 0.747]	2.594	0.010
movies	0.3750	[0.053, 0.697]	2.286	0.022
Number of articles viewed	**-0.00008508**	[-0.0000933, -0.0000769]	-20.339	< 0.01
(Intercept)	0.4794	[0.158, 0.801]	2.922	< 0.01

sity of articles viewed by users may not have been adequately represented. This could be due to the definition of GS-Score, which applies weighting based on the number of times the same article is read. However, the dataset used in this study did not record instances where a user read the same article multiple times, leading to an analysis under the condition that each article viewed was weighted uniformly. Additionally, there could be a bias in the vectors of the articles generated by the model used for the embedding of the news article.

Regarding the dataset used in this study, there were a total of 17 news article categories. In the multiple regression analysis exploring the relationship between article diversity and browsing categories, considering the impact of multicollinearity, we employed a stepwise method to use 14 significant categories as explanatory variables. However, there is still a possibility of multicollinearity among these 14 categories.

For future perspectives, improving the narrow distribution of GS-Scores could involve preparing a news dataset that considers the weight of each viewed article based on the number of views or viewing time, further selecting models for news article embedding, or creating models fine-tuned with Triplet Loss to distance article vectors from other categories. Additionally, to mitigate the impact of multicollinearity in the multiple regression analysis of the relationship between article diversity and browsing categories, evaluating the similarity and correlation between news article categories and considering categories with high similarity or correlation as the same could reduce the number of explanatory variables. Based on these two improvement strategies, we aim to reanalyze the relationship between the diversity of viewed articles and the tendencies in article viewing.

7 Conclusion

In this study, we utilized the natural language processing model Sentence-BERT to score the diversity of news articles viewed by users and conducted an investigation into the relationships and factors affecting the diversity of viewed articles. Specifically, we analyzed viewing tendencies from the aspects of the relationship between article diversity and sen-

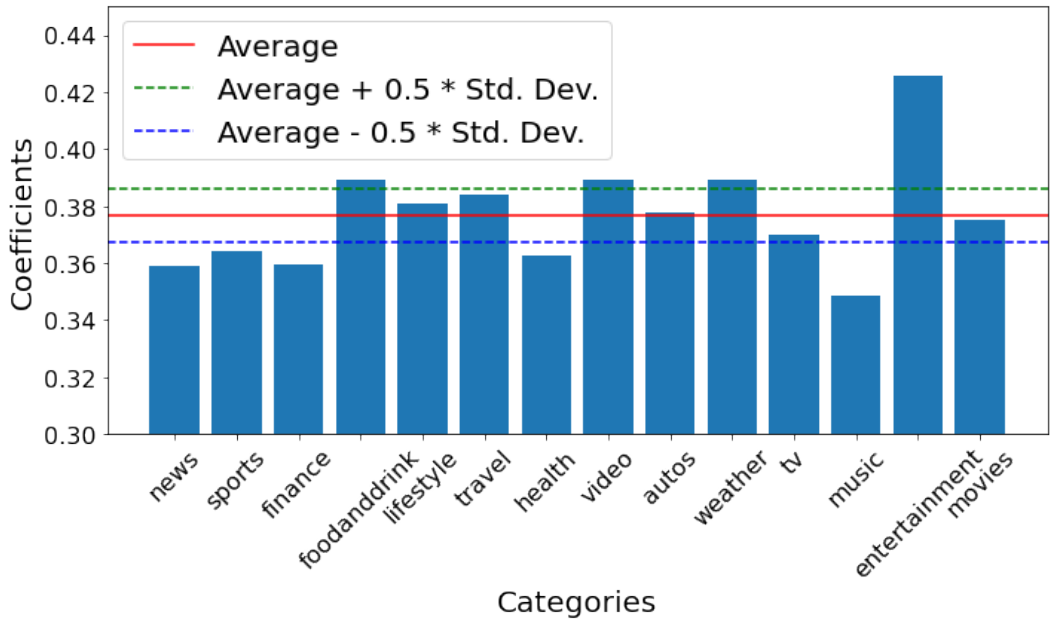

Figure 10: Coefficients of Each Category in the Multiple Regression Analysis

timents, as well as browsing category trends. For the embedding of news articles, we used paraphrase-MiniLM-L6-v2, a pretrained model of Sentence-BERT, and for the sentiment analysis model of news articles, we utilized a pretrained DistilBERT-based model, lxyuan/distilbert-base-multilingual-cased-sentiments-student.

Regarding the relationship between article diversity and sentiments, it was initially suggested that users tend to prefer negative news articles. This supports the findings of Robertson et al. [14]. The results of multiple regression analysis using the least squares method suggested that the more positive the users' viewing tendencies, the lower the diversity of viewed articles, or conversely, the more negative the users' viewing tendencies, the higher the diversity of viewed articles. Similarly, the analysis of the relationship between article diversity and browsing category trends revealed that news in categories such as music, current affairs, politics, crime incidents, world affairs, science, technology, and finance tend to have a low contribution to the diversity of information viewed. Conversely, categories like entertainment, food and dining, movies, and weather tend to have a high contribution to the diversity of information viewed. Furthermore, both analyses confirmed that users who view a greater number of articles tend to have higher diversity in their articles viewed, consistent with previous research [12].

However, this research is subject to several limitations. It is important to note that the GS-Score, used as an indicator of the diversity of users' articles in this analysis, showed a narrower distribution compared to previous research [12], suggesting that it may not fully represent the diversity of users' viewed articles. Additionally, consideration must be given to the potential insufficiency of the stepwise method alone in addressing the impact of multicollinearity in the multiple regression analysis of the relationship between article diversity and browsing categories. Furthermore, analysis similar to this study is needed to determine whether the findings apply to other news datasets.

Acknowledgements

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP23H00216 and the Cooperative Research Project Program of the Research Institute of Electrical Communication, Tohoku University, Japan.

References

[1] Ministry of Internal Affairs and Communications, "2023 White Paper on Information and Communications," Nikkei Printing Inc., 2023, pp. 30–31.

[2] C. R. SUNSTEIN, "republic: Divided democracy in the age of social media," Princeton University Press, 2018, pp. 98–136.

[3] B. Ytre-Arne and H. Moe, "Doomscrolling, monitoring and avoiding: News use in covid-19 pandemic lockdown," Journalism Studies, vol. 22, pp. 1–17, 2021.

[4] K. Ohata, K. Iizuka, and H. Yatomu, "Investigation of the impact of negative news on user behavior," The Database Society of Japan, 2023.

[5] F. Toriumi, and T. Yamamoto, "Towards a healthy discourse platform version 2.0 - implementing informational health," 2023; https://www.soumu.go.jp/main_content/000885478.pdf.

[6] M. Abdool, M. Haldar, P. Ramanathan, T. Sax, L. Zhang, A. Manaswala, L. Yang, B. Turnbull, Q. Zhang, and T. Legrand, "Managing diversity in airbnb search," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 2952–2960.

[7] A. Anderson, L. Maystre, I. Anderson, R. Mehrotra, and M. Lalmas, "Algorithmic effects on the diversity of consumption on spotify," in Proceedings of the Web Conference 2020, 2020, pp. 2155–2165.

[8] S. Suganuma, K. Iizuka, Y. Seki, and F. Toriumi, "Effect of Article Diversity on retention rates in Online News Service," Proceedings of the 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022, pp. 4H1OS2a03-4H1OS2a03; https://doi.org/10.11517/pjsai.JSAI2022.0_4H1OS2a03.

[9] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, and M. Zhou, "MIND: A large-scale dataset for news recommendation," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, July 2020, pp. 3597–3606; https://aclanthology.org/2020.acl-main.331.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[11] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019; https://doi.org/10.48550/arXiv.1908.10084.

[12] M. Sasaki, S. Okura, and S. Ono, "Analysis on the relationship between diversity of consumed news articles and user activity," Proceedings of the 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022, pp. 1H1GS1102–1H1GS1102; https://doi.org/10.11517/pjsai.JSAI2022.0_1H1GS1102.

[13] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan, "Exploring the filter bubble: the effect of using recommender systems on content diversity," Proceedings of the 23rd international conference on World wide web, 2014, pp. 677–686.

[14] C. Robertson, N. Pröllochs, K. Schwarzenegger, P. Pärnamets, J. Van Bavel, and S. Feuerriegel, "Negativity drives online news consumption," Nature Human Behaviour, vol. 7, Mar. 2023, pp. 1–11.

[15] I. Waller and A. Anderson, "Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms," The World Wide Web Conference, ser. WWW '19, New York, NY, USA: Association for Computing Machinery, 2019, pp. 1954–1964; https://dl.acm.org/doi/10.1145/3308558.3313729.

[16] D. Holtz, B. Carterette, P. Chandar, Z. Nazari, H. Cramer, and S. Aral, "The engagement-diversity connection: Evidence from a field experiment on spotify," Proceedings of the 21st ACM Conference on Economics and Computation, ser. EC '20, New York, NY, USA: Association for Computing Machinery, 2020, pp. 75–76; https://dl.acm.org/doi/10.1145/3391403.3399532.

[17] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," Lille, 2015.

[18] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation," in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, Eds. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 1–14; https://aclanthology.org/S17-2001.

[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[20] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, June 2015; https://ieeexplore.ieee.org/document/7298682.