electronics

MDPI

Article

Internet News User Analysis Using Deep Learning and Similarity Comparison

Sunoh Choi

Department of Software Engineering, Jeonbuk National University, Jeonju-si 54896, Korea; suno7@jbnu.ac.kr

Abstract: Nowadays, many Korean users read news from portal sites like Naver and Daum. Users can comment on news articles on such sites, and some try to influence public opinion through their comments. Therefore, news users need to be analyzed. This study proposes a deep learning method to classify each user's political stance. Further, a method is developed to evaluate how many similar comments each user writes, and another method is developed to evaluate the similarity of a user's comments with other users' comments. We collect approximately 2.68 million comments from hundreds of thousands of political news articles in April 2017. First, for the top 100 news users, we classify each user's political stance with 92.3% accuracy by using only 20% of data for deep learning training. Second, an evaluation of how many similar comments each user writes reveals that six users score more than 80 points. Third, an evaluation of the similarity of each user's comments to other users' comments reveals that 10 users score more than 80 points. Thus, based on this study, it is possible to detect malicious commenters, thereby enhancing comment systems used in news portal websites.

Keywords: internet news; deep learning; user analysis

Citation: Choi, S. Internet News User Analysis Using Deep Learning and Similarity Comparison. *Electronics* **2022**, *11*, 569. https://doi.org/10.3390/electronics11040569

Academic Editors: Nurul I. Sarkar and Juan-Carlos Cano

Received: 13 January 2022
Accepted: 11 February 2022
Published: 14 February 2022

1. Introduction

A survey by Korea's Ministry of Culture, Sports and Tourism found that 90% of people [1] read news from portal sites like Naver [2] or Daum [3]. Specifically, as smartphone use increases, people are likelier to read the news from portal sites than from offline or online newspapers. In this light, the selection of news on portal sites is attracting increasing interest [4].

People can comment on the news, and some try to influence public opinion through their comments. In 2012, some staff at Korea's national information service tried to influence public opinion by writing comments on news portal sites [5]. In 2017, a user nicknamed Druking tried to influence the public opinion by writing comments using the Kingcrap system that writes many similar comments using several smartphones [6]. In 2022, all political parties tried to influence public opinion by writing comments on news portal sites [7].

Therefore, news users who write comments on news portal sites must be studied. Some users try to influence public opinion by repeatedly writing similar comments; others write comments similar to other users' comments. This study aims to analyze users to accurately understand the public opinion without being affected by malicious comments.

For this purpose, we collected approximately 2.68 million comments written by 200,000 users on 100,000 political news articles on Daum [2] in April 2017. News comments from April 2017 were collected instead of current news comments from 2022 to avoid political controversy; notably, a presidential election was held in May 2017.

We propose three methods to analyze users. First, we classify each user's political stance by using the Seq2Seq deep learning model [8]. Training using only 20% of all data resulted in a classification accuracy of 92.3%. Second, we analyzed how many similar comments each user writes. We found that among the top 100 users, six scored above

80 points in this regard. Third, we analyzed the similarity of each user's comments with other users' comments. We found that 10 users scored above 80 points in this regard.

The remainder of this paper is organized as follows. Section 2 presents related works. Section 3 describes the three proposed methods to analyze users. Section 4 presents the experimental results of these three methods. Finally, Section 5 presents conclusions of this study.

2. Related Work

Wikipedia [9] provides information about various topics. However, this information may sometimes contain biases that must be removed. Recasens et al. discussed the framing bias and epistemological bias and identified common linguistic cues for them [10]. Hube et al. proposed a supervised classification approach that depends on an automatically created lexicon of bias words [11].

Fan et al. investigated the effects of information bias, that is, factual content that is presented to influence readers' opinions [12]. Cho et al. proposed a method to classify the political bias of news articles using subword tokenization [13]. In our study, we classified users' political stances and proposed methods to evaluate how many similar comments each user writes and how similar these comments are to other users' comments.

Recently, Korea's conservative party launched the Kraken artificial intelligence system [7] to detect malicious comments on news portal sites. This system is similar to ours in detecting malicious comments; however, its algorithm has not been presented.

Garrett suggested that the desire for opinion reinforcement may play an essential role in shaping individuals' exposure to the political information provided online [14]. The results demonstrated that opinion-reinforcing information promotes the exposure to news stories, whereas opinion-challenging information makes such exposure less likely. The objective of this study is different from that of Garrett's study, as we analyze news portal users based on their comments.

Koroniotis et al. conducted a study on determining abnormal activities in an Internet of Things (IoT)-based network [15]. They proposed the particle swarm optimization technique to obtain hyperparameter values in deep neural networks. The objective of their study is different from that of our study.

Ming et al. conducted a study on identifying a person having a same identity from several cameras [16]. Because they focused on image processing using a deep learning-based approach, the objective of their study differed from that of our study.

Yao et al. conducted a study on the advantages of using deep learning in IoT systems [17]. They applied DeepSense to user identification through biometric motion analysis (UserID). Therefore, the objective of their study differed from that of our study; however, there exists one similarity between the two, as both studies use deep learning to solve their respective research problems.

3. Internet News User Analysis Method

3.1. News and Comment Data

We collected approximately 100,000 news articles from Daum in April 2017. Then, we collected approximately 2.68 million comments written by 200,000 users on these articles. We created the database tables News_list and Comments to respectively store news data and comment data.

News_list (Num, Subject, Post_ID, Company, News_Time, News_Date)

Comments (Num, ID, Count, Content, Time, Post_ID, Name, Company)

Table 1 lists the top 5 news articles in terms of number of comments. The top news article had 9754 comments written by 7427 users, indicating that each user wrote 1.3 comments (i.e., more than one) on average.

Table 1. Top 5 news articles in terms of number of comments.

Num	Post_ID	Subject	Company	Num of Comments	Num of Users
1	20170407085950355	Ahn Copies Obama's Speech	Herald	9754	7427
2	20170407163335851	Sewol-ho wasting 0.1B$	Yonhap	9266	8067
3	20170402160552920	Taegukgi Gathering in Bongha	News 1	7367	6484
4	20170401154147438	Moon, Park Amnesty	Newsys	6687	4917
5	20170404111730189	Stray TK	Edaily	5553	4765

Table 2 lists user data for the top 5 commenters. The first user wrote 673 comments on 634 news articles in one month; in other words, this user wrote an average of 23 comments each day. Further, the second user wrote 652 comments on 250 news articles; in other words, this user wrote an average of 2.6 comments for each news article.

Table 2. User data for top 5 commenters.

Num	ID	Name	Num of Comments	Num of News Articles
1	9010433	Sleeping User	673	634
2	-135404000	Happycat	652	250
3	16476611	Kwakyongwoo	643	486
4	-144133664	Candy	622	461
5	-118722416	Moonse~	599	276

3.2. Deep Learning Model to Classify News User's Political Stance

We proposed a deep learning model based on Seq2Seq to classify a user's political stance [8], as shown in Figure 1. In the first step, we collected news and comment data. In the second step, we labeled the political stance for the top 100 users. In the third step, we extracted words from the comments. In the fourth step, we classified each user's political stance using the Seq2Seq model. We used only 20% of the data for training the deep learning model. The classification accuracy is discussed in Section 4.

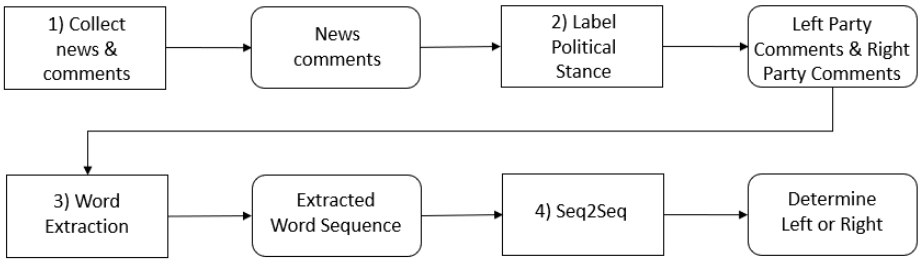

Figure 1. Process to classify each user's political stance.

Figure 2 shows the Seq2Seq deep learning model used to classify each user's political stance. This model consists of an embedding layer and a long short-term memory (LSTM) layer [18]. The model's parameters are listed in Table 3.

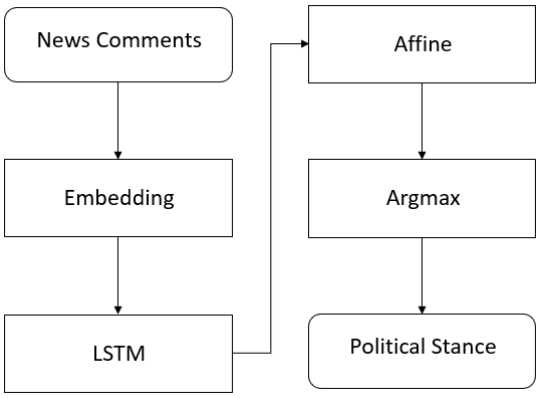

Figure 2. Seq2Seq model to classify each user's political stance.

Table 3. Parameters of the proposed deep learning model.

Parameter	Value
vocab_size	3443
wordvec_size	8
hidden_size	16
batch_size	10
max_epoch	100

3.3. Method to Evaluate How Many Similar Comments Each User Writes

We proposed a method to evaluate how many similar comments each user writes. As noted in Section 3.1, each user wrote several comments for each news article.

We used the Jaccard similarity [19,20] to evaluate how similar two comments are to each other. We extracted two word sets $S_{i,j,1}$ and $S_{i,j,2}$ from two comments $C_{i,j,1}$ and $C_{i,j,2}$ for news article N_i and user U_j. Then, the similarity between the two comments was calculated as

$$Sim\left(C_{i,j,1}, C_{i,j,2}\right) = \frac{S_{i,j,1} \cap S_{i,j,2}}{S_{i,j,1} \cup S_{i,j,2}} \times 100$$

As shown in Figure 3, the user U_j may write l comments for news article N_i. The similarity score $SimScore\left(N_i, U_j\right)$ was calculated as:

$$SimScore\left(N_i, U_j\right) = \sum_{k=2}^{l} Sim(C_{i,j,1}, C_{i,j,k}) / (l-1)$$

In addition, the user U_j may write comments for n news articles. The similarity score was calculated as:

$$SimScore\left(U_j\right) = \sum_{i=1}^{n} SimScore\left(N_i, U_j\right) / n$$

By using the similarity score for each user, we evaluated how many similar comments each user wrote.

3.4. Method to Evaluate Similarity of Each User's Comments to Other Users' Comments

Next, we proposed a method to evaluate the similarity of each user's comments to other users' comments. As shown in Kingcrap [7] in 2017, a malicious user can write similar comments from several IDs.

We used the Jaccard similarity [19,20] to evaluate the similarity of one user's comments to another user's comments as shown in Figure 4. We extract word set $S_{i,p,1}$ from comment

$C_{i,p,1}$ from user U_p on news article N_i and word set $S_{i,q,1}$ from comment $C_{i,q,1}$ from user U_q on news article N_i. The similarity score between the two users' comments is calculated as:

$$SimBtwUsers\left(C_{i,p,1}, C_{i,q,1}\right) = \frac{S_{i,p,1} \cap S_{i,q,2}}{S_{i,p,1} \cup S_{i,q,1}} \times 100$$

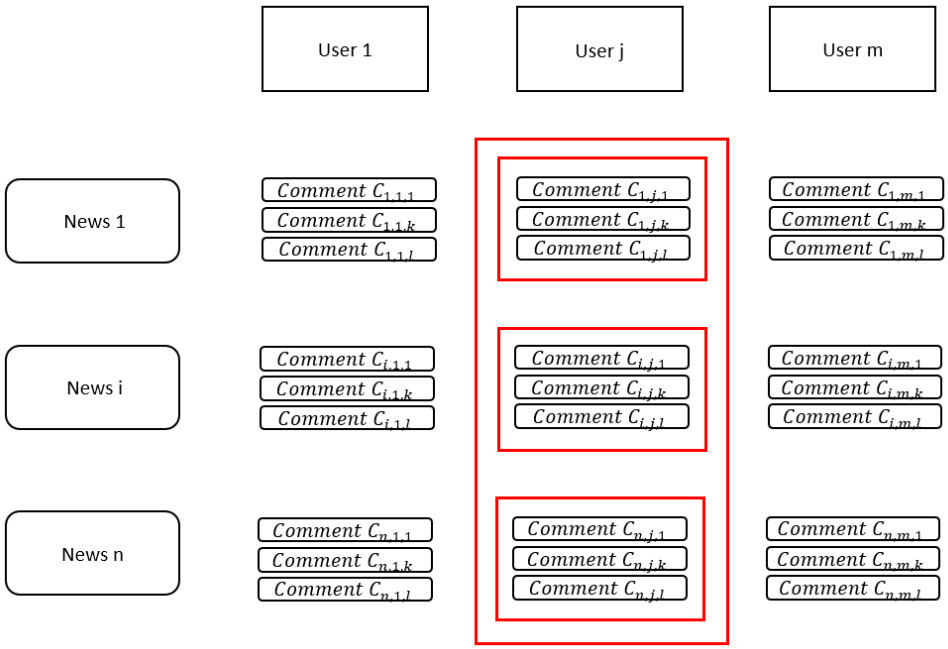

Figure 3. Method to evaluate how many similar comments each user writes.

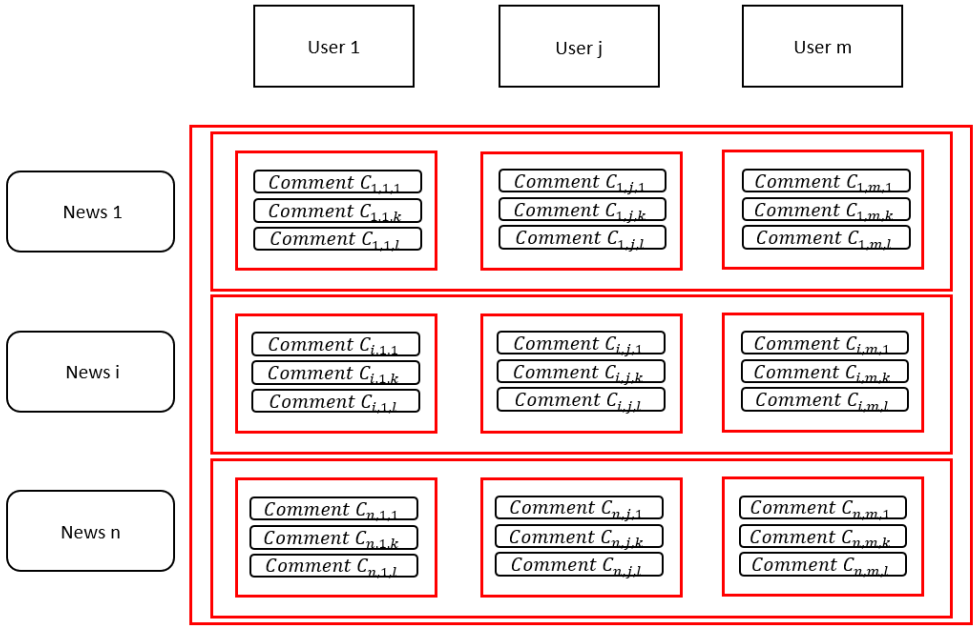

Figure 4. Method to evaluate similarity of each user's comments to other users' comments.

m users comment on news article N_i. Therefore, we should evaluate the similarity of user U_i's comments to other users' comments. The similarity score between users for news article N_i is calculated as:

$$SimScoreBtwUsers(N_i, U_p) = Max(SimBtwUsers(C_{i,p,1}, C_{i,j,1}))$$

By using the similarity score between users, we can evaluate the similarity of each user's comments to other users' comments.

Finally, we calculate the similarity score of user U_p for all news articles. This similarity is calculated as:

$$SimScoreBtwUsers(U_p) = Max(SimScoreBtwUsers(N_i, U_p))$$

4. Experimental Results

4.1. Setup

The experimental environment is as follows. We used a computer with an Intel i7 3.7 GHz CPU and 16 GB of memory that was running Windows 10. We used Python 3.7 to collect news and comment data and a MySQL database to store them. To parse comments, we used BeautifulSoup4 [21]. In addition, to extract Korean words, we used Hannanum [22]. Finally, we used Keras 2.0 [23] for the deep learning framework.

4.2. Analyzing News User's Political Stance using Deep Learning

First, we performed an experiment to classify each user's political stance using Seq2Seq [8]. We used the top 100 users' comment data. For training and testing, we used 20 and 80 users' comment data, respectively. We performed five-fold cross validation. For comparison, we used 1, 5, and 10 comments per user. As shown in Figure 5, when we used 1, 5, and 10 comments per user, the accuracy was 87.65%, 67.65%, and 92.4%, respectively.

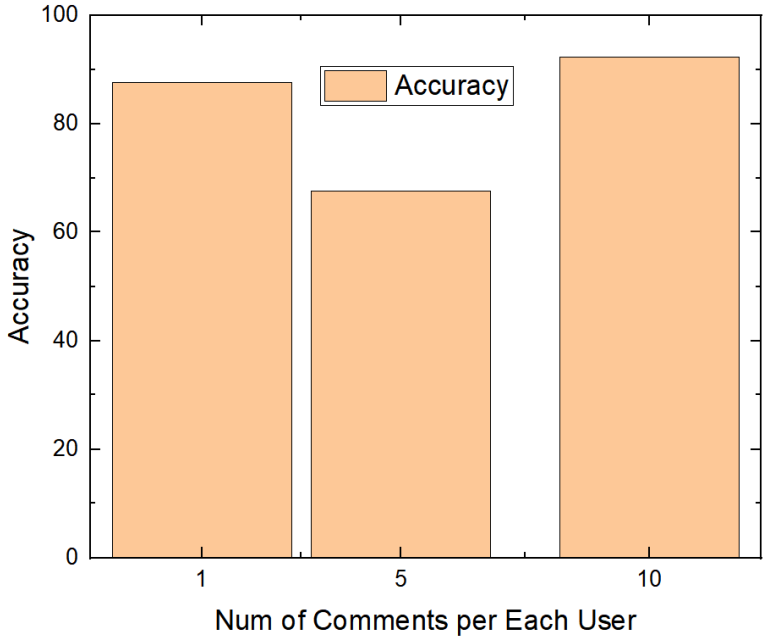

Figure 5. Accuracy of classifying each user's political stance using Seq2Seq.

The use of one comment per user was not enough to classify each user's political stance. When using five comments per user, the accuracy reduced compared to that when using one comment. However, when using 10 comments per user, the accuracy was increased by 3.69% compared to that when using one comment. This suggests that a suitable number of

comments needs to be used for each user to classify their political stance. Notably, only 20% of the data were used for training. When 80% of data were used for training as usual, the accuracy was 100%.

To effectively classify the users' political stances, we only used the top 100 users. Therefore, the basic Seq2Seq model was sufficient for this analysis. However, in our future studies, to analyze the political stances of additional users, we would require a more precise deep learning-based model.

4.3. Analyzing News Users' Comments by Similarity Comparison Method

Second, we evaluated how many similar comments each user wrote using the similarity comparison method. As shown in Figure 6, 69 users had scores lower than 20; 14 users had scores between 20 and 40; 7 users had scores between 40 and 60; 4 users had scores between 60 and 80; and 6 users had scores higher than 80. Note that when a user wrote the same comments, the score was 100.

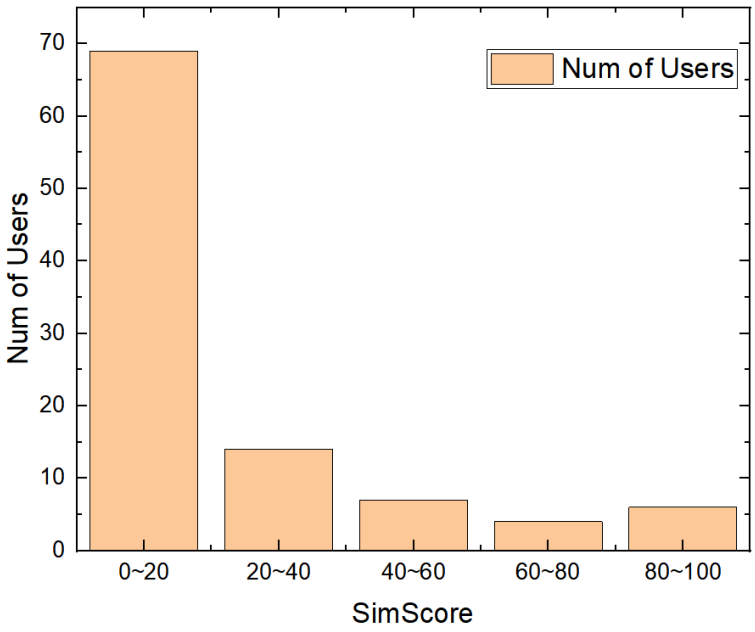

Figure 6. Comparison of comment similarity.

Table 4 lists users whose scores are greater than 80.

Table 4. Top 6 users by similarity score.

Num	ID	Name	Score
1	-128438191	명상의교훈 (Lesson of Meditation)	95.11
2	-135404000	happycat	93.22
3	-72027861	제노비오 (Zenovio)	91.54
4	-109241458	RICHMAN	91.33
5	-69404078	사춘기 (Puberty)	91.28
6	-127763077	코알라똥구멍 (Koala~)	85.01

Table 5 lists the comments of the user having the highest similarity score. It shows that this user wrote the same comments for the same news article every 30 s.

Table 5. Comments of user having the highest similarity score.

Num	ID	Name	Post_ID	Time
1	-128438191	Lesson of Meditation	20170407103227790	2017-04-07T10:58:14+0900
		These guys from student movement are below the level.		
2	-128438191	Lesson of Meditation	20170407103227790	2017-04-07T10:49:50+0900
		These guys from student movement are below the level		
3	-128438191	Lesson of Meditation	20170407103227790	2017-04-07T10:49:03+0900
		These guys from student movement are below the level		
4	-128438191	Lesson of Meditation	20170407103227790	2017-04-07T10:48:44+0900
		These guys from student movement are below the level		
5	-128438191	Lesson of Meditation	20170407103227790	2017-04-07T10:48:28+0900
		These guys from student movement are below the level		

4.4. Analyzing News Users Using Similarity Comparison

Finally, we evaluated the similarity of each user's comments to other users' comments, as shown in Figure 7. The results indicated that 52 users had scores lower than 20; 31 users had scores between 20 and 40; 3 users had scores between 40 and 60; 4 users had scores between 60 and 80; and 10 users had scores higher than 80. Note that when a user wrote the same comment as another user, the score was 100.

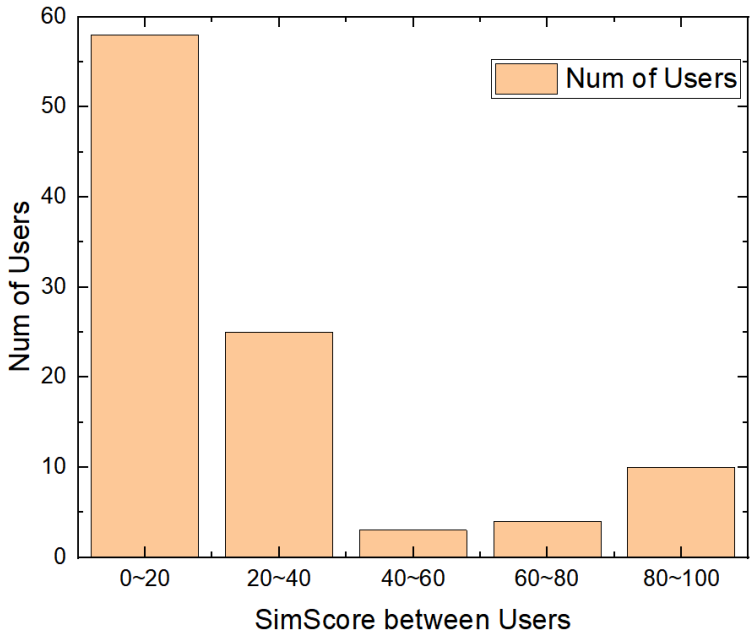

Figure 7. Similarity score between users.

Table 6 lists users with scores higher than 80.

Table 6. Top 10 users by similarity score.

Num	ID_1	Name_1	ID_2	Name_2	SimScoreBtwUsers
1	-135404000	Happycat	-72027861	Zenovio	100
2	-144133664	Candy	-40781432	Audrey~	100
3	-72027861	Zenovio	-109241458	RICHMAN	100
4	-109241458	RICHMAN	-72027861	Zenovio	100
5	-116992362	EULJI~	-107059556	Goguryeo	100
6	-2723829	~Moon	-124844642	Roro	100
7	-107059556	Goguryeo	-116992362	Eulji~	100
8	-40781432	Audrey~	-144133664	Candy	100
9	-124844642	Roro	-2723829	~Moon	100
10	-103273764	Gwanggae~	-116992362	Eulji~	87.5

Table 6 indicates that four groups wrote similar comments, as shown in Figure 8.

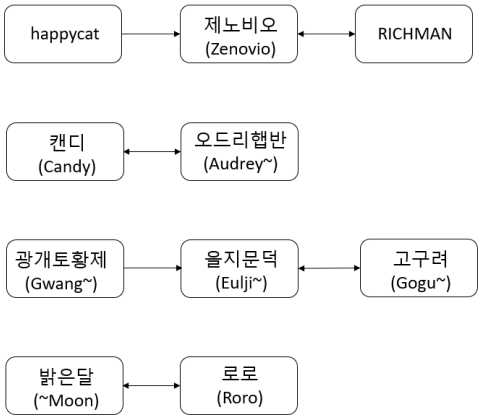

Figure 8. Groups who write similar comments.

Table 7 lists the top 2 users' comments. These two users wrote the same comments for the same news articles.

Table 7. Top 2 users' comments.

Num	ID	Name	Post_ID	Time
1	-135404000	happycat	20170411100831	2017-04-11T12:28:09+0900
		Moon is not ready candidate		
2	-72027861	Zenovio	20170411100831	2017-04-11T10:49:57+0900
		Moon is not ready candidate		

We used the Jaccard similarity for comparing the comment similarity. By evaluating the similarity of each user's comments to other users' comments, we showed that several groups wrote similar comments. In future work, we aim to use graph neural networks [20,24] to find additional information about the relations between users.

Finally, we gave the top users a political stance, as depicted in Figure 9. The six users presented on the right side of the figure wrote similar comments, as listed in Table 4, while the three groups presented in Figure 8 are depicted on the left side of Figure 9.

Figure 9. Top users' political stance.

Overall, we collected approximately 2.68 million comments on hundreds of thousands of news articles. However, owing to a lack of computing power, we analyzed only the top 100 users and top 100 news articles. In future work, we aim to reduce the computational complexity to analyze a larger number of users.

Additionally, we only collected comments uploaded over a period of one month. In our future studies, we plan to collect and analyze comments uploaded over a period of at least five years from 2017 to 2022 to determine the change in the number of users.

5. Conclusions

In Korea, many Internet users read news from portal sites. On such sites, users can comment on news articles, but some users attempt to influence public opinion through their comments. In this study, we analyzed such users of news portal sites.

To achieve the aforementioned objective, we proposed three methods for analyzing the users of news portal websites. First, we developed a deep learning method based on the Seq2Seq model to classify each user's political stance [8]. Subsequently, we developed a method for evaluating the number of similar comments written by each user. Finally, we developed a method for evaluating the similarity between each user's comments and other users' comments.

For the top 100 news portal site users, we first classified each user's political stance and achieved an accuracy level of 92.3%. Next, we evaluated the number of similar comments written by each user, and the results revealed that six users scored over 80 points. Finally, we evaluated the similarity between each user's comments and other users' comments, and the results revealed that 10 users scored over 80 points. Hence, based on these results, we can conclude that it is possible to enhance the performance of comment systems used in news portal sites, especially with regard to the way in which such systems can be used for the detection of malicious commenters.

Funding: This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2019R1G1A11100261) and was supported by research funds for newly appointed professors of Jeonbuk National University in 2021.

Conflicts of Interest: Authors declare no conflict of interest.

References

1. User Ratio Reading News from Portal Sites. Available online: https://www.dailyimpact.co.kr/news/articleView.html?idxno=50488 (accessed on 10 January 2022).
2. Naver. Available online: http://www.naver.com (accessed on 10 January 2022).
3. Daum. Available online: http://www.daum.net (accessed on 10 January 2022).
4. Choi, D. Internet Portal Competition and Economic Incentive to Tailor News Slant. *Korean J. Ind. Organ.* **2017**, *25*, 40.
5. Ji-Hye, J. Assembly's NIS Prove Fizzles out, KoreaTimes. 2018. Available online: http://www.koreatimes.co.kr/www/nation/2013/08/113_141397.html (accessed on 10 January 2022).
6. Suh-yoon, L. Governor Kim Kyoung-soo Sentenced to 2 Years for Online Opinion Rigging, KoreaTimes. 2019. Available online: http://www.koreatimes.co.kr/www/nation/2019/01/113_262961.html (accessed on 10 January 2022).

7. Shin, H. Kraken to Detect Malicious Comments, JoongAng. 2021. Available online: https://www.joongang.co.kr/article/25036975#home (accessed on 10 January 2022).
8. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. *arXiv* **2014**, arXiv:1409.3215.
9. Wikipedia. Available online: https://en.wikipedia.org/wiki/Main_Page (accessed on 10 January 2022).
10. Recasens, M.; Danescu-Niculescu-Mizil, C.; Jurafsky, D. Linguistic Models for Analyzing and Detecting Biased Language. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; pp. 1650–1659.
11. Hube, C.; Fetahu, B. Detecting Biased Statements in Wikipedia. In Proceedings of the World Wide Web Conference, Lyon, France, 23–27 April 2018.
12. Fan, L.; White, M.; Sharma, E.; Su, R.; Choubey, P.K.; Huang, R.; Wang, L. In Plain Sight: Media Bias through the Lens of Factual Reporting. In Proceedings of the 2019 Conference on Empirical Methods in National Language Processing, Hong Kong, China, 3–7 November 2019; pp. 6343–6349.
13. Cho, D.B.; Lee, H.Y.; Jung, W.S.; Kang, S.S. Automatic Classification and Vocabulary Analysis of Political Bias in News Articles by Using Subword Tokenization. *KIPS Trans. Softw. Data Eng.* **2021**, *10*, 1–8.
14. Garrett, R.K. Echo chambers online?: Politically motivated selective exposure among Internet news users. *J. Comput.-Mediat. Commun.* **2009**, *14*, 265–285. [CrossRef]
15. Koroniotis, N.; Moustafa, N.; Sitnikova, E. A new network forensic framework based on deep learning for Internet of Things networks: A particle deep framework. *Future Gener. Comput. Syst.* **2020**, *110*, 91–106. [CrossRef]
16. Ming, Z.; Zhu, M.; Wang, X.; Zhu, J.; Cheng, J.; Gao, C.; Yang, Y.; Wei, Z. Deep Learning-based person re-dientification methods: A survey and outlook of recent works. *Image Vis. Comput.* 2022; *in press*. [CrossRef]
17. Yao, S.; Zhao, Y.; Zhang, A.; Hu, S.; Shao, H.; Zhang, C.; Su, L.; Abdelzaher, T. Deep Learning for the Internet of Things. *IEEE Comput. Mag.* **2018**, *51*, 32–41. [CrossRef]
18. Olah, C. Understanding LSTM Networks. Available online: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (accessed on 10 January 2022).
19. Jaccard Index. Available online: https://deepai.org/machine-learning-glossary-and-terms/jaccard-index (accessed on 10 January 2022).
20. Choi, S. Malicious Powershell Detection using Graph Convolution Network. *Appl. Sci.* **2021**, *11*, 6429. [CrossRef]
21. Beautifulsoup4. Available online: https://pypi.org/project/beautifulsoup4/ (accessed on 10 January 2022).
22. Hannanum. Available online: https://konlpy-ko.readthedocs.io/ko/v0.4.3/api/konlpy.tag/ (accessed on 10 January 2022).
23. Keras. Available online: https://keras.io/ (accessed on 10 January 2022).
24. Kipf, N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.