Journal of Universal Computer Science, vol. 31, no. 7 (2025), 735-755
submitted: 28/5/2024, accepted: 14/4/2025, appeared: 28/6/2025 CC BY 4.0

Examining Deep Learning Techniques for Ethical Artificial Intelligence: Cleansing Malicious Comments from Users

Ji Woong Yoo
(Syntekabio, Daejeon, Republic of Korea
https://orcid.org/0009-0001-5897-4962, jwyoo@syntekabio.com)

Kyoung Jun Lee
(Kyung Hee University, Seoul, Republic of Korea
https://orcid.org/0000-0002-7774-6551, klee@khu.ac.kr)

Arum Park
(Kwangwoon University, Seoul, Republic of Korea
https://orcid.org/0009-0000-9557-328X, penellope007@kw.ac.kr)

Abstract: The advancement of AI has heightened the significance of ethical concerns, particularly in managing negative user feedback like malicious comments, necessitating thoughtful deliberation. The focus of this research is to explore the potential of deep learning techniques in addressing these issues and enhancing the ethical nature of AI systems. Specifically, we investigated the collection and processing of news comment data using Long Short-Term Memory (LSTM) algorithm and Word2Vec model. The primary objective was to evaluate how deep learning techniques can improve the quality of data obtained from news comments. Our findings demonstrate that deep learning models surpass CleanBot in accuracy and block rates for handling negative user comments, including malicious ones, enabling organizations to effectively manage such comments in online communities using AI-based methods. This study adds to the existing research by showing how advanced deep learning techniques can effectively identify and classify harmful comments by analyzing complex language patterns.

Keywords: Artificial intelligence, Deep learning, Malicious data, Media studies, Ethical data, Content filtering, User generated content, LSTM, Word2Vec
Categories: I.2, I.7, K.4, E
DOI: 10.3897/jucs.128450

1. Introduction

AI pioneers from the 1950s aspired to construct robots capable of sensing, reasoning, and thinking like humans. In recent years, AI has become increasingly important across numerous sectors [Daugherty & Wilson, 2018]. As AI develops, its ethical implications have become a major issue [Brendel et al., 2021; Farisco et al., 2020; Johnson & Verdicchio, 2023; Siau & Wang, 2020]. The development of ethically sound AI systems necessitates a focus on the quality of data used to train these systems. Modern advances in computing and the availability of large datasets enable AI technologies to go beyond simple automation and information retrieval. However, it is crucial to ensure

the quality of big data to develop AI that is both useful and socially acceptable. Additionally, considering ethical aspects is essential for responsible AI development [Muller et al., 2021]. Furthermore, societal issues and the growing demand for information have shed light on the importance of AI systems learning factors beyond traditional data analysis. For instance, AI should now be capable of selecting news articles that captivate readers' interest, providing personalized article recommendations [Zhang et al., 2021], and predicting reactions and attitudes towards media content that may elicit controversy or sensitive responses [Matalon et al., 2021]. This expansion in AI capabilities is exemplified by machine learning algorithms that detect suspicious financial transactions and recommend fraud management decisions [Perols, 2011], as well as smart bots and vehicles autonomously delivering essential supplies like food and medicine [Chung, 2021].

The field of media and communication studies has traditionally focused on exploring interpersonal communication, neglecting a comprehensive examination of various dimensions related to AI [Guzman & Lewis, 2020]. As AI is poised to influence users' media consumption choices and their evaluations of news and social media content, media communication scholars must investigate AI's ethical advancement. AI learning is important. Current AI systems are trained on large datasets. AI can learn from the many editorial and journalistic-reviewed news pieces. Conversely, what about user reactions or comment data, which represent another substantial corpus of information generated through public engagement? While user-generated content may supplement news stories or captivate the interest of other content consumers, it may also be marred by profanity, bias, and unethical conduct such as racism [Zou & Schiebinger, 2018]. The challenge lies in the sheer volume of this type of data, which cannot be disregarded since AI necessitates a comprehensive understanding of human responses to make informed judgments about information.

Artificial intelligence (AI) is gaining significance in modern businesses, offering opportunities for automation, data analysis, and customer engagement [Siau & Wang, 2020]. Its capability to process large volumes of data swiftly and accurately has become crucial for maintaining competitiveness in the present-day market [Huang et al., 2022]. Additionally, as AI continues to advance in its capacity to understand and interpret human language, it becomes an increasingly valuable tool in customer service and sales. AI integrates into our everyday lives through virtual assistants and smart home devices, simplifying tasks and improving overall effectiveness [Siau & Wang, 2020]. Popular virtual assistants such as Apple's Siri, Amazon's Alexa, and various language learning model-based applications are increasingly enhancing their functionalities beyond basic tasks like checking the weather and managing calendars. They now include customer service and learning support features. AI possesses the capability to autonomously generate and select decisions, which can occasionally lead to ethical violations according to human standards [Huang et al., 2022]. Moreover, AI may approach ethical judgments differently than humans, as its standards are based on mathematical calculations rather than subjective moral beliefs [Johnson & Verdicchio, 2023]. Therefore, while AI offers efficient solutions, it is essential to consider the ethical implications of granting AI the power to make decisions that impact human life and well-being [Farisco et al., 2020]. To date, the ethical aspects of training AI systems have predominantly focused on limited applications such as spam filtering [Alberto et al., 2015] and mitigating malicious comments [Baccouche et al., 2020]. While these endeavors have provided valuable contributions, there is a need to expand the scope

and direction of AI research in addressing ethical concerns. In this context, our research aims to explore how a deep learning approach, specifically leveraging convolutional neural networks (CNNs), gated recurrent units (GRUs), and attention mechanisms, can enhance the quality of data in news commentaries.

The central question driving this study is how deep learning techniques can be effectively utilized to enhance the quality of news commentary data generated by users. Deep learning techniques, known for their ability to identify complex patterns and contextual information within textual data, are particularly useful for detecting malicious user remarks. Unlike traditional rule-based approaches, which may struggle to handle the subtleties and variations in language that characterize such remarks, deep learning techniques offer a more robust and nuanced approach [Munir et al., 2018; P. Wang et al., 2021]. To address this research question, we conducted data collection and validation from a prominent Korean news website. We performed a comparative analysis, evaluating the performance of the news portal's existing malicious comment filtering technique against suggested models incorporating CNNs, GRUs, and attention mechanisms. The results of our study highlight the potential benefits of employing this integrated approach to enhance the quality of news commentary data.

User-generated content has become a significant source of big data generation since the advent of Web 2.0, which emphasizes user participation and collaboration on online platforms. The vast amounts of text data provided by users through blogs, social media, and other platforms are rich with features, including individual words, combinations of words, and their order, all of which contribute to understanding context and meaning. Traditional machine learning algorithms often struggle with capturing these nuanced linguistic characteristics due to their simpler feature sets. To address this issue, recent advancements in natural language processing have emerged, employing techniques such as deep learning and neural networks. Deep learning techniques have proven to be highly effective in filtering out misinformation and hate speech from user-generated content.

This research contributes to the current discourse on AI utilization from a social-technological perspective, emphasizing the increasing importance of AI applications in the present context. This paper aims to shed light on the crucial role of data quality in training ethically driven AI systems and presents a novel approach to improving the quality of news commentaries using deep learning techniques. The implications of this study are expected to be significant in the current era where the utilization of AI from a socio-technological perspective is rapidly gaining prominence.

2. Background

2.1. Ethical Artificial Intelligence

Artificial intelligence (AI) is becoming an increasingly important aspect of modern businesses, providing numerous opportunities for automation, data analysis, and customer engagement [Siau & Wang, 2020]. With its ability to quickly and accurately process large amounts of data, AI has become a vital tool for staying competitive in today's market [Huang et al., 2022]. Its diverse range of applications continues to expand, from chatbots [e.g., Luo et al., 2019] to predictive analytics [e.g., Agrawal et al., 2018], making it an indispensable presence in organizations across various

industries [Brendel et al., 2021]. Moreover, as AI's ability to comprehend and interpret human language improves, it becomes an increasingly valuable instrument for customer service and sales [Siau & Wang, 2020]. AI is seamlessly incorporated into our daily lives through virtual assistants and smart home devices, making tasks simpler and more efficient. Even while there are still concerns about how artificial intelligence can affect employment, there is reason to be optimistic about the potential of the technology to enhance human labor by opening up new avenues for cooperation and increasing levels of output [Johnson & Verdicchio, 2023].

AI can make decisions on behalf of humans because it can generate multiple options and autonomously select the best option. However, this ability can sometimes result in decisions that violate ethical standards set by humans [Huang et al., 2022]. Additionally, AI may defer ethical judgments in a way that humans may not, and the ethical standards humans hold may differ from those that AI operates on, which are based on mathematical calculations rather than subjective moral beliefs [Johnson & Verdicchio, 2023]. Therefore, while AI can provide efficient and effective solutions, it is important to consider the ethical implications of allowing AI to make decisions that could affect human life and well-being.

Due to their constant evolution, the opaque nature of artificial intelligence (AI) algorithms poses a challenge for both comprehending and establishing norms for their use [McCorduck & Cfe, 2004]. This issue is also exacerbated by the difficulty of comprehending and validating AI decisions [Felzmann et al., 2019]. Even for the creators, the lack of transparency, commonly referred to as the 'black box' in the application of AI algorithms, poses distinct ethical challenges [W. Wang & Siau, 2019]. As the influence of AI on society continues to grow, there are various topics, such as autonomous driving, self-directed weapon systems, and cockpit automation, that raise societal considerations and even ethical dilemmas related to matters of life and death [Brendel et al., 2021]. To mitigate the potential ethical concerns and societal repercussions associated with artificial intelligence (AI), organizations must take a proactive approach to addressing these issues [Dinah et al., 1997]. While some may view AI as an indispensable tool in the technology arms race, organizations must assume responsibility for considering the ethical implications of AI [Makridakis, 2017]. From AI development to deployment, all stages of AI utilization must be scrutinized to circumvent the risk of unethical and harmful consequences. In particular, the use of AI in relation to customers requires careful consideration to ensure that AI-based technologies are not being utilized in a way that could lead to significant societal harm [Brendel et al., 2021; Huang et al., 2022; Siau & Wang, 2020]. By being proactive in addressing these issues, organizations can ensure that AI is used ethically and responsibly, benefiting both themselves and society as a whole [Farisco et al., 2020].

According to Johnson and Verdicchio (2023), it is argued that directly applying human ethical standards when judging ethical AI is difficult. Their argument is that AI is nothing more than a machine that processes complex, mathematical questions based on algorithms and computational capabilities. The concern is that while human ethical judgment is sometimes relative, flexible, and social in nature, AI may have ethical standards that are more rigid, less heuristic, and more distant from social concerns. The development of ethical artificial intelligence necessitates giving ethical data a higher priority and implementing an ethical development procedure. Additionally, it is crucial to involve various stakeholders to enhance the ethicality of AI development processes. Establishing an independent and democratically accountable advisory body is

necessary. Academic efforts must be made to develop internationally accepted ethical guidelines, supported by legal and political initiatives. In addition, it is essential to ensure that AI is developed in a way that allows for continuous human oversight to ensure ethical outcomes. It is problematic and a result of human ignorance to fail to properly develop and monitor ethical AI.

It is worth noting that AI designed for human communication will learn from data on various topics that are currently prevalent in society [Chen & Zimbra, 2010]. The use of biased user opinions or posts to train AI is likely to raise ethical concerns [Farisco et al., 2020]. In particular, the opinions of a limited number of people can be highly biased, and the anonymous nature of the Internet makes it difficult to determine whether short user-generated comments meet ethical standards [Singer & Ashman, 2009, Buccafurri et al., 2022]. Huang et al. (2022) point out that it is difficult to predict ethical flaws in AI when the quality of the data is poor, so it is important to improve the quality of the data required for AI training. Efforts to ensure that AI is developed and operated in a way that refines its data to learn human ethical standards, controls itself in context, and discusses ethical issues with humans are becoming increasingly essential [Huang et al., 2022].

2.2. Abusive comments in online

While there is some overlap in their definitions, online trolling and cyberbullying are distinct phenomena. Online trolling refers to the act of intentionally posting inflammatory or offensive comments with the goal of provoking a reaction from others. Cyberbullying, on the other hand, involves the use of technology to harass, intimidate, or humiliate someone repeatedly over time. Posting malicious or hateful comments on content posted by others on social media, or persistently engaging in malicious behavior to harass a user, are all unethical and reprehensible behaviors [Modha et al., 2020]. Both online trolling and cyberbullying can cause severe mental discomfort, anxiety, sadness, and, in some cases, suicide among their victims [López-Meneses et al., 2020]. In addition to the psychological harm caused by online trolling and cyberbullying, victims may also experience social consequences, such as a loss of trust in others and a reduced willingness to engage with others online. Establishing appropriate response systems for victims of such actions on online platforms and implementing measures to prevent recurrence or proactively prevent these behaviors are essential for fostering a healthy online community [Brown et al., 2006].

Social media, such as Twitter, Facebook, and Instagram, are online platforms for sharing personal posts and opinions. Users get satisfaction from sharing information or exchanging messages with other users with whom they have a weak connection, and the platforms earn revenue by providing these services, such as through advertising or game sales [Kane et al., 2014]. Social media connections do not require offline intimacy, and in many cases, individuals can be connected indirectly or through public activity, which may expose personal information to anonymous parties, leaving them vulnerable to aggressive and malicious messages [Bowler et al., 2015]. Cyberbullying is not limited to social media platforms like Twitter and Facebook. It can occur on any internet platform with user engagement, including news comment sections [Song & Song, 2021], YouTube comments [Kyriacou & Zuin, 2016], online food delivery platforms [Barnes, 2018], and more.

Online virtual spaces are often viewed as a 'forum' where individuals can express their opinions more freely [Wattal et al., 2010]. It can be difficult to distinguish between a person's outspoken opinions and malicious intent, particularly on topics where political or philosophical differences are inevitable [Fromm et al., 2018]. For example, many Twitter messages written during politically sensitive times like the presidential election were sometimes biased, offensive, and filled with factual inaccuracies [Allcott & Gentzkow, 2017; Bovet & Makse, 2019; Brummette et al., 2018]. While trolling or cyberbullying can be determined when there is a clear intent to attack a specific individual on social media by including racial or sexually degrading language, or profanity or hateful words, it is not necessarily possible to make a mechanical and inflexible determination in all cases [Malmasi & Zampieri, 2018]. Distinguishing between offensive language in a post about a socially sensitive event or expressing empathy for an individual's unfortunate incident, and what constitutes trolling and cyberbullying, can be challenging [Aro, 2016; Peebles, 2014]. It should be noted that determining what constitutes ethical expression requires careful consideration of context, which involves a complex decision-making process. Cultural or social contexts can determine the justification level of ethical expression, but it is important to recognize that this can be relative and sometimes biased [Fromm et al., 2018].

Malmasi and Zampieri (2018) proposed a machine learning approach using deep learning models to distinguish between general profanity and hate speech in social media. Experiment results have shown that it is not easy to distinguish between those two concepts based on the superficial composition of words, thus requiring further research efforts. Corazza et al. (2020) proposed a method for detecting hate speech in multilingual environments. Modha et al. (2020) have proposed an algorithm that categorizes offensive language into overtly aggressive, covertly aggressive, and non-aggressive labels, making them visually distinct on social media. By incorporating psychological factors, Balakrishnan et al. (2020) showed that a more effective method for determining cyberbullying messages on Twitter could be developed. These research efforts are largely driven by the need to develop artificial intelligence algorithms to select messages quickly and automatically to be filtered on social media, either by detecting words containing profanity or swearing, or by utilizing feedback data from other users.

The possibility that AI learning data may contain biases and result in undesirable outcomes has necessitated the development of sophisticated filtering techniques. There has been limited research on the impact of AI on the news media, which plays a significant role in shaping public opinion and promoting social change, even though its impact on social media has received considerable attention. Given the high level of social engagement associated with news media, this dearth of research is particularly concerning. To ensure that AI is used ethically and effectively, it is necessary to expand research efforts beyond social media and invest in the development of sophisticated media filtering technologies. Online news articles often feature user-engaged comments, which can sometimes lead to malicious reactions to sensitive information, resulting in extreme consequences such as celebrity suicides or unfounded prejudice against a region [Kim, 2022; Nam, 2019]. As AI requires extensive and concurrent training data, news and its associated comments and messages, as they circulate through social media, can all be utilized. Therefore, it is crucial to consider filtering techniques for news comments as an important aspect of developing more ethical AI.

3. Methodology

3.1. Data

Abusive (or malicious) comments have become a major issue in Korean society, particularly in cases of sensitive social issues or political conflicts, as well as news articles about celebrities where commenting is allowed. In Korea, the dissemination of information via the internet is remarkably rapid. People are geographically and culturally close, and the scope of freedom of expression online is quite broad. However, equally frequent are the spread of fake news, attacks on celebrities, and political figures through social media. The freedom of expression that the internet provides has led to an increase in individuals expressing their opinions, sometimes abusively. Unfortunately, this has also led to a significant increase in cases where individuals have suffered both emotional and financial damages due to malicious comments. With similar cases repeatedly occurring, examining the negative aspects of abusive comments is not particularly difficult.

On October 14, 2019, Sully, a Korean actress, was found dead on the second floor of her home. Her sudden death shocked many people. She publicly spoke out in favor of women's rights, and it was reportedly very distressing for her to see unrelated malicious comments on news stories about her. It once banned users from commenting on entertainment articles on Daum, a large South Korean online portal. Kim Gun-mo, a popular South Korean singer, was subjected to constant malicious comments from Kang Yong-seok, a Korean YouTube influencer, in December 2019, alleging that Kim committed sexual assault against several women, until prosecutors and courts cleared Kim of the charges in November 2022. Unverified information was spread through online news, and abusive and insulting comments were posted on the articles of those suspicious news sources, which tended to spread the story further. Eventually, during this period, Kim had to shut down his economic activities and focus on countering fake news and malicious comments. On the one hand, political news stories are frequently accompanied by a variety of malevolent comments. It is a debatable point, as the process of expressing political opinions in comments can result in heated discussions, not all of which are necessarily detrimental to molding public opinion.

In this study, we focused 360 articles and 91,716 comments about celebrities and politicians on Naver, the largest news portal in South Korea, from November 12 to December 14, 2019, and crawled the comments on those articles for analysis. This was the time when the investigation into Kim Gun-mo's sexual assault was underway and slanderous articles about the President Moon Jae-in began to circulate. We chose a crucial time for both celebrities and politicians, when very heated discussions were taking place across news and comments in Korea. We chose South Korea as our primary source of data for two reasons. First, news articles and comments have played a very important role in shaping public opinion in South Korea. For example, South Korean politician Kim Kyung-soo tried to gain political advantage by manipulating comments, which was the main reason he was sent to prison. Second, in South Korea, various information registered on news portals, including comments, can potentially be utilized for AI development. Since this importance is related to the empirical and practical utility of this research, this study focuses on a specific country and a specific point in time, which creates unique strengths and potential limitations that will be discussed later.

Subject	Number of comments	Number of blocked comments	Block rate (%)
Moon Jae-in	24,726	1,102	4.46
Kim Gun-mo	66,990	1,809	2.70
Total	91,716	2,911	3.17

Table 1: Number of comments and bloc rate by Naver Clean Bot

The data was gathered through a Python-based web-crawler and Selenium testing tools. We ensured that the data collection process did not impose a burden on Naver and was solely intended for research purposes. The dataset we collected did not contain any sensitive information that could potentially compromise privacy or cause harm to individuals. At the time of data collection, Naver already categorized and blocked abusive or malicious comments with a service called 'Clean Bot.' Naver has its own dictionary of profanity and tries to protect the rights of users as much as possible. Table 1 shows the data and the blocking rate of malicious comments by Clean Bot.

3.2. Deep learning approaches

Traditional machine learning techniques such as SVM and Naive Bayes primarily operate based on structured features like word frequency or TF-IDF, which limits their ability to capture complex contextual information. While RNN or LSTM-based models are proficient in handling sequence data, they may be less efficient than GRUs in processing long-term dependencies. Furthermore, excluding CNNs and the Attention Mechanism may result in an inadequate reflection of local features or the importance of words within sentences. In conclusion, integrating CNN, GRU, and the Attention Mechanism combines the strengths of different mechanisms to handle the subtleties and variations of language more accurately and effectively. This integration enhances model performance, providing higher accuracy and better contextual understanding compared to existing technologies.

3.2.1. Word embedding

In natural language processing (NLP) problems, such as text categorization, the problem of converting words into vectors arises. In this case, one-hot encoding is a vector representation method using one-hot vectors, where the value of the vector corresponding to the position of the target is one (=1) and all other indices are represented as zero (=0). However, this method has the disadvantage that most of the values in the matrix are sparse vectors represented by zeros, and the similarity between words cannot be calculated. A way to solve this problem is Word2Vec, which can compute similarities between words and represent them as dense vectors (Aggarwal, 2018). Word2Vec is primarily computed using a bag of words (CBOW) or skip-gram, a model that predicts the surrounding words from the center word, and its structure is made up of a single silver layer of neural networks (Jang et al., 2019). In this study, we need to identify the characteristics of a specific word through the occurrence of surrounding words, so we implemented Word2Vec through the skip-gram method. In this case, the algorithm trained to maximize the probability P(c|o) that a peripheral word o appears given a central word c. In addition, since the computational burden of Word2Vec increases rapidly as the number of words and hidden layers increases, to reduce the amount of training for frequently occurring corpora, we consider the

frequency of words and proceed with learning whenever a less frequent word appears. In other words, the probability of learning a word, PWi=1-tf(wi), is determined by the percentage of occurrence of a word, f(wi), and the learning parameter, t.

3.2.2. Recurrent neural network

Traditional deep neural networks (or deep learning) have the disadvantage of not being able to account for data with continuous characteristics. A recurrent neural network (RNN) is a model designed to solve problems with such continuous data [Goodfellow, 2017]. Comments on news articles are organized in sequences of words, which is relevant to the need to adopt the RNN approach [Z. Wang et al., 2020]. Sequential data refers to data where each piece of data is in a sequence, such as audio signals, text, and stock price data. The RNN is characterized by the fact that each type of step t takes as input the input vector Xt and the output vector ht-1 of the previous step, so it uses the weight Wx of each node xt and the weight Wh of ht-1, respectively such as, ht=Ø(WxTxt+WhTht-1+b).

RNN models use internal state (memory) to process variable length sequences of inputs, which makes them ideal for tasks such as language modeling and time-series analysis; however, basic RNNs suffer from the vanishing gradient problem, which makes it hard for them to learn and model long-term dependencies in the data [Ribeiro et al., 2020]. The gated recurrent unit (GRU), one of the solutions to this problem, is a type of RNN used in deep learning [Zulqarnain et al., 2019]. It improves upon the basic RNN by introducing gating units. The GRU has two distinct gates: (1) update gate (z): This gate determines the extent of past information (from previous time steps) that should be carried forward to the future. It synthesizes a blend of new input and past memory. If the value of the update gate is near 1, it allows the past memory to pass through, whereas if it's near 0, it permits new input to pass through. (2) reset gate (r): This gate determines the amount of past information to discard. If the reset gate's value is near 0, it instructs the hidden state to disregard past memory and reset with the current input. These gates equip the GRU with the ability to adaptively capture dependencies spanning different time scales. Consequently, the network can opt to forget the stored state and directly accept the new input for short-term dependencies or lean more heavily on the historical state for long-term dependencies.

3.2.3. Attention mechanism

Sequence-to-sequence (Seq2seq) models are utilized in deep learning for tasks such as natural language processing, machine translation, and speech recognition because they receive a sequence of data as input and output a sequence of data [Kuznetsov & Mariet, 2019]. Since a sentence can be thought of as a sequence of words, Seq2seq can be used for AI development, such as chatbots and translators that show input and output in the form of sentences [Sutskever et al., 2014]. Seq2seq models typically utilize two different architectures: an encoder and a decoder. The encoder converts the input sequence data into a compact vector, which is then passed to the decoder to create the sequence data. The hidden state ht of the encoder cell is determined by the input xt and the hidden state ht-1 of the previous step as follows: ht=f(xt,ht-1). The context vector c is the hidden state of the last cell of the encoder, which is determined by the hidden state of all previous steps as follows: c=qh1,h2,⋯,hTx. Thus, the probability of word presence is, py=t=1Tp(yt|y1,y2,⋯,yt-1,c).

In Seq2Seq models, the compact vector representation has a fixed length. When dealing with long sequences, it is anticipated that the model may encounter performance issues due to information loss. To address this, it becomes crucial to employ a method that ensures the encoder pays sufficient attention to all hidden states, even for longer sequences. This can be achieved through techniques such as attention mechanisms, which enable the model to effectively capture the relevant information from the entire sequence, mitigating the performance penalty associated with longer sequences. Thus, the conditional probability will follow a separate compact vector ci for the predicted word yi as follows: si=f(si-1,yi-1,ci), and pyiy1,⋯,yi-1,x=g(yi-1,si,ci). The dynamic nature of attention enables the model to accommodate varying sequence lengths and align source and target sequences more precisely, resulting in enhanced translation quality, text summarization, and other sequence-to-sequence tasks.

4. Design

To utilize text data as input, it is necessary to convert words into vector representations. In this particular study, the researchers opted for syllable-based embeddings instead of word-based embeddings. The decision was driven by the understanding that word-based embeddings might not be suitable for news comment data due to the presence of various expressions such as typos, non-standard expressions, and Korean vowels. Furthermore, when creating word sets using syllables as tokens, syllables with a frequency of 4 or fewer occurrences were deemed as rare syllables and subsequently removed from the word set. In the comment data analyzed, rare syllables accounted for approximately 29.5% in the word set and 4.6% in the complete dataset. This approach aimed to refine the word set by excluding infrequent syllables, thus ensuring a more focused and representative dataset for analysis.

In this study, we design a model to determine malicious comments and text classification is considered as the task of automatically organizing them according to predefined classes or topics. According to the portal's policy, the maximum number of characters for comments on Naver's newspapers is 300, and most comments are within 150 characters in length (see Figure 1).

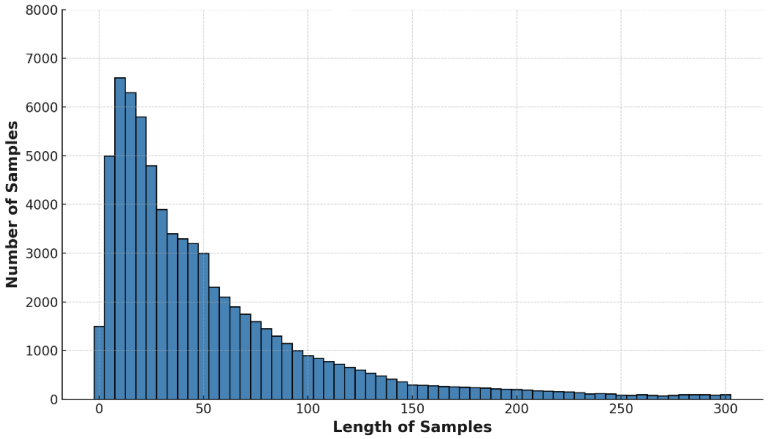


Figure 1: Length distribution of sentence data

To assess the impact of incorporating various methods that consider shared context, sentence arrangement, and contextual information on performance, we developed a model depicted in Figure 2. This model combines techniques such as Gated Recurrent Units (GRUs), attention mechanisms, and Convolutional Neural Networks (CNNs) to capture the intricacies of the data and optimize performance. By integrating these approaches, we aimed to leverage the strengths of each method and create a comprehensive model capable of achieving enhanced results.

Textual data is transformed into embeddings syllable by syllable. We collected news comments containing typos, grammatical problems, and unusual characters. To overcome these issues, preprocessing is needed. We chose Python to address these underlying issues. Subsequently, the Korean text is disassembled into individual syllables, thereby mitigating the vulnerability to typos and misspellings. This method has been praised for managing Hangul, the Korean writing system [Choi, 2000].

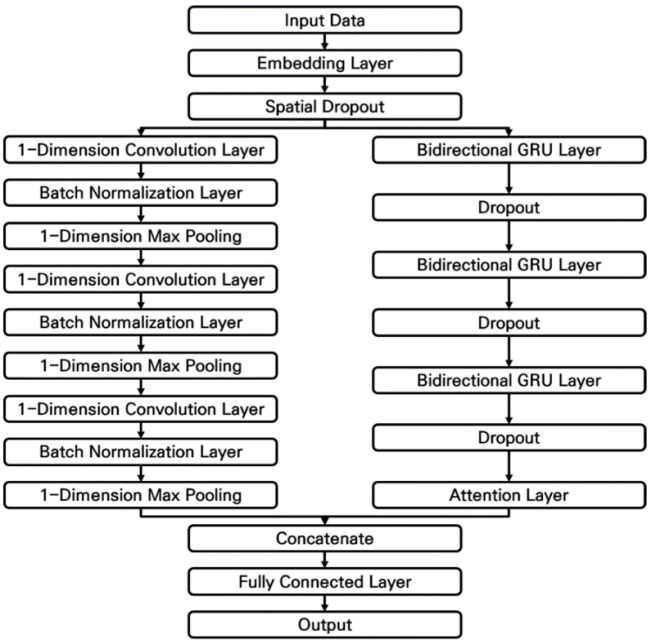


Figure 2: Deep learning processes

Due to their shorter lengths and irregularities, an initial convolutional layer classified harmful remarks. Convolutional neural networks (CNNs) were chosen because they can abstract non-linear patterns. Embedding data represented each corpus as a dense vector, which was converted from two-dimensional to one-dimensional before being input into the deep learning model. This conversion lost spatial structure information. The convolutional layer, which accepted two-dimensional input, retained data shape and connected only with data within its weights' filters. After capturing low-level information, following convolutional layers abstracted high-level abstract qualities. In this study, batch normalization was applied to three convolutional layers

to improve training efficiency. Convolutional neural networks (CNNs) have traditionally been used in computer vision [Goodfellow, 2017]; however, since text classification models that combine word embeddings and convolutional neural networks have shown superior performance, many attempts have been made to utilize convolutional neural networks for NLP [Yin et al., 2017]. For instance, Zhang et al. (2015) conducted a study where they discovered that a CNN-based NLP model demonstrated strong performance in text classification tasks that involved character embeddings.

Sequential input dependencies make the convolutional neural network (CNN) technique unsuitable for text classification [Yin et al., 2017]. For this, recurrent neural networks (RNNs) are more appropriate, with the gated recurrent unit (GRU) providing comparable performance to the long short-term memory (LSTM) while offering faster learning [Van Huynh et al., 2019]. Bidirectional GRUs capture inter-word interactions by evaluating both prior and succeeding sentence contexts [Sun et al., 2019]. In this study, three bidirectional GRUs with dropout were employed for text classification. Attention strategies let GRU frameworks avoid learning irrelevant words. This study's self-attention passes the final GRU cell's secret state, revealing the sequence's influential spots. Self-referential self-attention may improve contextual knowledge in natural language processing tasks.

5. Result

5.1. Labelling comments

Three evaluators classified comments to build a performance comparison dataset. Inter-rater reliability (IRR) was determined to assess evaluation reliability [Chaturvedi & Shweta, 2015]. Table 2 displays IRR. Fleiss's Kappa is a statistical measure of rater agreement when categorizing items into numerous groups (Banerjee et al., 1999). Fleiss's Kappa adjusts for chance agreement. It accurately assesses agreement among several raters or observers by accounting for chance agreement. Kappa values vary from -1 to 1, indicating rater agreement. 0 implies probability agreement, whereas 1 shows absolute agreement. A score below 0 indicates worse agreement than expected by chance.

Category	Kappa (std error)	Z (p-value)	Conditional probability
Not malicious	0.485 (0.002)	248.72(<0.001)	0.920
Malicious	0.485 (0.002)	248.72(<0.001)	0.565
Overall	0.485 (0.002)	248.72(<0.001)	-

Table 2: Inter-rater reliability for labelling user comments

The Pearson correlation coefficient measures linear correlation and changing patterns in two populations. This study examined labeled data discriminant validity using a correlation analysis (Table 3). The IRR analysis and inter-rater correlation coefficient are identical.

	Rater A	Rater B	Rater C
Rater A	1		
Rater B	0.565***	1	
Rater C	0.474***	0.502***	1
*: p-value < 0.05, **: p-value < 0.01, ***: p-value < 0.001			

Table 3: Correlations between rating scores

5.2. Performance comparison

This study compared CleanBot, a benchmark instance, to suggested models. Table 4 shows that the convolutional neural network (CNN), gated recurrent unit (GRU), and attention model had the highest accuracy. The same methodology also blocked the most harmful remarks (see Table 5). The CNN model performed best for profanity-, sexual-, and disparaging terms that use irregularly changing profanity (Figure 3). The GRU-attention mechanism model performed better for violent- and discriminating phrases, which require sentence context. The CNN+GRU+Attention model performed best.

	Accuracy (%)	Blockage rate (%)
CleanBot (benchmark)	55.87	11.87
Alternative A	87.01	79.88
Alternative B	87.12	79.01
Alternative A + B	88.21	84.39
* Alternative A: CNN; Alternative B: GRU + Attention		

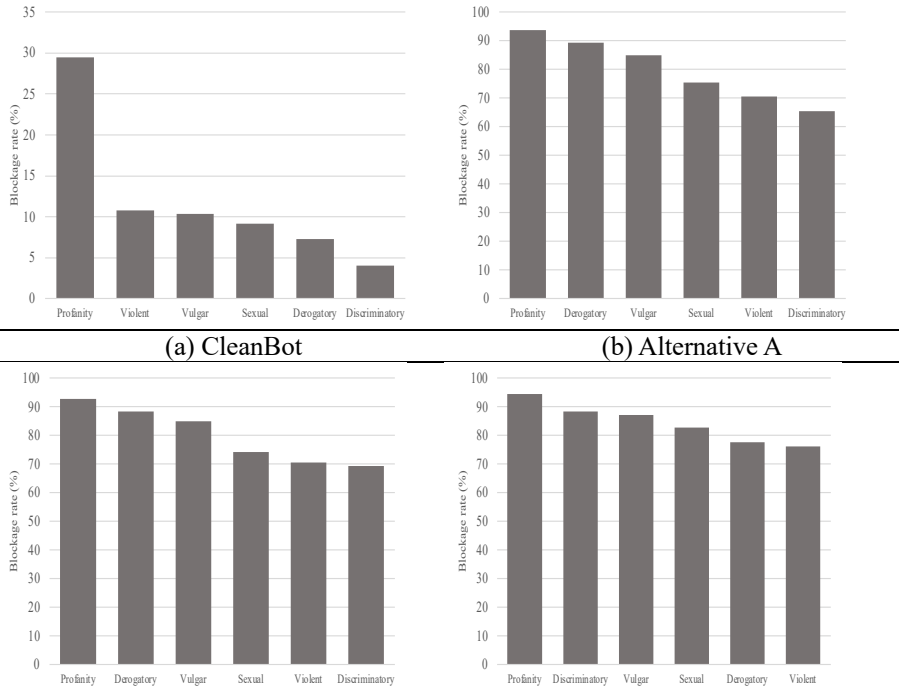
Table 4: Blockage rates for each approach

CleanBot is used to increase user participation without limiting it. Thus, it blocks more content when dangerous user comments contain profanity. CleanBot blocks fewer violent, profane, sexual, and derogatory statements. Note that CleanBot's blocking rate for significant harmful expressions does not necessarily indicate a critical viewpoint or functional issue in its user policy. From AI training data, CleanBot may not be polished enough to manage user feedback. If social media businesses want to give users freedom of speech and flexibility in comment tone, they will need to update their training data for writing comments with artificial intelligence.

		Blockage rate by expression category (%)				
	Profanity	Vulgar	Sexual	Violent	Discriminatory	Derogatory
CleanBot	29.44	10.38	9.18	10.79	4.08	7.32
Alternative A	93.59	85.01	75.51	70.50	65.31	89.37
Alternative B	92.76	85.01	64.28	74.10	69.39	88.44
Alternative A + B	94.41	87.51	82.65	76.26	77.55	88.29
* Alternative A: CNN; Alternative B: GRU + Attention						

Table 5: Blockage rates by for each category and approach

Deep learning models often remove profanity. The CNN-based Alternative A approach blocked fewer violent and racist statements. These results imply that the GRU + Attention algorithm may mitigate violent and discriminating utterances better than CleanBot and Alternative A.



(a) CleanBot (b) Alternative A

(c) Alternative B (d) Alternative A + B

* Alternative A: CNN; Alternative B: GRU + Attention

Figure 3: Blockage rates by for each category and approach

5.3. Related work

S. Wang et al. (2014) highlights the challenge of recommending trustworthy web services and proposes a novel reputation measurement approach. Web service suggestions are improved by recognizing fraudulent feedback and moderating subjective preferences. Real-data experiments indicate reduced reputation measurement variance and improved suggestion success ratio. Our study filters dangerous content using user-generated data in artificial intelligence, unlike others. Our study also takes a novel approach to screening dangerous content by using user participation to create data and deep learning. These contributions show how our research can advance content filtering and user-generated data analysis.

Alberto et al. (2015) discusses the challenges of comment spam on YouTube due to the platform's limited moderation tools. Spam has forced popular channel owners to block comments. The author compares decision trees, logistic regression, Bernoulli

Naive Bayes, random forests, linear and Gaussian SVMs to solve this problem. Building upon this analysis, Alberto et al. (2015) introduces TubeSpam, an online system designed to accurately filter comments on YouTube. While our study aligns with Alberto et al. (2015) discussion on the importance of user comments, there is a distinction in the context. News stories and user-generated content are textual and excellent learning targets for AI language models. Current news and user replies are important training data for AI, according to our research. This distinction highlights our study's focus on AI training data.

Baccouche et al. (2020) conducted a study with the objective of identifying spam emails, spam comments, and fraudulent articles. For preprocessed text classification, they used deep learning and LSTM networks. Their analysis combined social media and fake email data in an innovative way. Their research suggested using a multi-label LSTM model to identify fraudulent text. Distinguishing our study from the work conducted by Baccouche et al. (2020), we focused on user comments to news articles, highlighting their dissemination through social media and their inherent value in terms of information sharing. Our research takes a unique approach because such remarks are the main cause of user reactions. Incorporating GRU and Attention processes deepened our analysis. These findings highlight our study's new approach of training AI systems with user input.

Campolo and Crawford (2020) say deep learning techniques are popular in AI because they can find patterns in large datasets and make accurate classifications and predictions. They also show the lack of theoretical explanations for these approaches' exceptional success. To address this gap, Campolo and Crawford (2020) draw upon Max Weber's theory of disenchantment with emphasizing the importance of human control and monitoring. Contrasting with the study conducted by Campolo and Crawford (2020), our research offers practical and empirical ways to support their findings. Our study stresses data preprocessing in ethical AI development. Our research sheds light on ethical issues and suggests ways to create ethical AI systems.

The topics discussed by Bhatt (2019) highlight the increasing shift of society towards a data-centric paradigm, emphasizing the growing significance of trust in data. He underlines the need of collecting and using large volumes of social media data and its reliability for providing services like AI assistants. In comparison to the study by Bhatt (2019), our research highlights user participation by using user-generated comment data for AI training. Our study analyzes user reactions to highly important news pieces, emphasizing its relevance and importance. The results of this study also add new perspectives on using user-generated data in AI systems, particularly in news analysis and user reactions.

Finally, ethical considerations must be heeded when monitoring user-generated news content with deep learning. Bowler et al. (2015) highlight that deep learning models can inherit and amplify biases present in the training data, leading to potentially unfair moderation practices that may disproportionately affect certain groups. Additionally, deep learning models are often considered black boxes due to their complexity, making it difficult to explain their decisions regarding parameter weights. Ensuring transparency in moderation practices and monitoring results is crucial for maintaining user trust and accountability. Determining accountability for decisions made by AI systems can be challenging, necessitating the establishment of clear accountability frameworks to ensure responsible AI deployment. Future research should focus on developing robust accountability mechanisms that delineate

responsibility and provide clear guidelines for the deployment and oversight of AI systems.

6. Conclusions

This study addressed internet abusive comments. We created CNN, GRU, and attention mechanism deep learning models by evaluating news items and user comments from a popular South Korean news portal. We wanted to promote ethical AI content filtering and harmful comment detection. Our models outperformed CleanBot, a comment filtering system, in blocking various sorts of fraudulent remarks. The combination of CNN, GRU, and attention mechanism model outperformed the others, demonstrating the importance of shared context, phrase order, and contextual information in filtering offensive remarks.

This research advances AI and content filtering. First, we used word embedding, RNNs, and attention algorithms to recognize and categorize malicious comments. These methods captured complicated linguistic patterns and context for accurate categorization in our models. Second, we compared several methods, including CleanBot, for abusive comment filtering. Our proposed model had better accuracy and blockage rates.

The research has real applications for content screening AI systems, demonstrating improved accuracy in blocking malicious comments and enhancing abusive content detection and filtering. This has practical consequences for internet platforms seeking safer and more courteous settings. The models showcased the impact of context in comment classification using RNNs and attention processes, aiding in the creation of AI systems that consider user interactions beyond keyword-based screening. The study emphasized the relevance of user involvement in AI model construction and refining, highlighting that platforms can improve content filtering by leveraging user-generated data.

This study contributes to the development of AI content filtering, but it is important to note that there are certain limitations associated with it. The investigation centered on analyzing news portal articles and user comments from a specific timeframe, which may have restricted the scope of the findings. The study examined the ethical implications of AI without specifically investigating the biases and impact on freedom of expression caused by content filtering. This topic necessitates additional research. While the models exhibited accuracy, fairness, and transparency, further development is required, along with thorough testing and validation, prior to implementation. Although this research serves as a valuable foundation for further exploration into the ethical aspects of AI, it fails to examine the potential policy consequences of implementing ethical principles in AI development. Subsequent investigations should adopt a comprehensive and interdisciplinary methodology, taking into account social and cultural factors.

Despite these limitations, the research lays the groundwork for responsible AI content screening and online abuse models. Establishing ethical development procedures in AI requires involving various stakeholders to enhance the ethicality of AI processes. An independent and democratically accountable advisory body is essential, and academic efforts must develop internationally accepted ethical guidelines supported by legal and political initiatives. As AI utilization increases, these

discussions must occur immediately and continue sustainably. The conclusion has been revised based on the inspiration drawn from the reviewer's question, with sincere thanks for their thoughtful engagement. Additionally, developing measurements to evaluate the ethical extent of user-generated comments could be an interesting topic. Establishing more accurate and fair measurements can proactively contribute to improving the quality of AI training data.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5A8080527)

References

[Aggarwal, 18] Aggarwal, C. C.: Neural networks and deep learning: a textbook. Yorktown Heights, N.Y.: Springer, 2018. http://snu-primo.hosted.exlibrisgroup.com/82SNU:82SNU_INST21620464370002591 http://snu-primo.hosted.exlibrisgroup.com/82SNU:82SNU_INST51624382420002591

[Agrawal, 18] Agrawal, A., Gans, J., Goldfarb, A.: Prediction machines: the simple economics of artificial intelligence. Harvard Business Press, 2018.

[Alberto, 15] Alberto, T. C., Lochter, J. V., Almeida, T. A.: Tubespam: Comment spam filtering on YouTube. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015.

[Allcott, 17] Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211-236, 2017. https://doi.org/10.1257/jep.31.2.211

[Aro, 16] Aro, J.: The cyberspace war: propaganda and trolling as warfare tools. European View, 15(1), 121-132, 2016. https://doi.org/10.1007/s12290-016-0395-5

[Baccouche, 20] Baccouche, A., Ahmed, S., Sierra-Sosa, D., Elmaghraby, A.: Malicious text identification: deep learning from public comments and emails. Information, 11(6), 1-19, 2020. https://doi.org/10.3390/info11060312

[Balakrishnan, 20] Balakrishnan, V., Khan, S., Arabnia, H. R.: Improving cyberbullying detection using Twitter users' psychological features and machine learning. Computers & Security, 90, 101710, 2020. https://doi.org/10.1016/j.cose.2019.101710

[Banerjee, 99] Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D.: Beyond kappa: A review of interrater agreement measures. Canadian Journal of Statistics, 27(1), 3-23, 1999. https://doi.org/10.2307/3315487

[Barnes, 18] Barnes, W. R.: The good, the bad, and the ugly of online reviews: the trouble with trolls and a role for contract law after the Consumer Review Fairness Act. Georgia Law Review, 53(2), 549-612, 2018.

[Bhatt, 19] Bhatt, J. J.: I Think I Can Trust Alexa, But How Much? Intersect: The Stanford Journal of Science, Technology, and Society, 13(1), 2019.

[Bovet, 19] Bovet, A., Makse, H. A.: Influence of fake news in Twitter during the 2016 US presidential election. Nature Communications, 10, 1-14, 2019. https://doi.org/10.1038/s41467-018-07761-2

[Bowler, 15] Bowler, L., Knobel, C., Mattern, E.: From cyberbullying to well-being: A narrative-based participatory approach to values-oriented design for social media. Journal of the Association for Information Science and Technology, 66(6), 1274-1293, 2015. https://doi.org/10.1002/asi.23270

[Brendel, 21] Brendel, A. B., Mirbabaie, M., Lembcke, T.-B., Hofeditz, L.: Ethical management of artificial intelligence. Sustainability, 13(4), 2021. https://doi.org/10.3390/su13041974

[Brummette, 18] Brummette, J., DiStaso, M., Vafeiadis, M., Messner, M.: Read all about it: The politicization of "fake news" on Twitter. Journalism & Mass Communication Quarterly, 95(2), 497-517, 2018. https://doi.org/10.1177/1077699018769906

[Buccafurri, 22] Buccafurri, F., De Angelis, V., Idone, M. F., & Labrini, C.: A protocol for anonymous short communications in social networks and its application to proximity-based services. Online Social Networks and Media, 31, 2022. https://doi.org/10.1016/j.osnem.2022.100221

[Campolo, 20] Campolo, A., Crawford, K.: Enchanted determinism: Power without responsibility in artificial intelligence. Engaging Science, Technology, and Society, 6, 1-19, 2020. https://doi.org/10.17351/ests2020.277

[Chaturvedi, 15] Chaturvedi, S., & Shweta, R.: Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods. Journal of the Indian Academy of Applied Psychology, 41(3), 20-27, 2015.

[Chen, 10] Chen, H., & Zimbra, D.: AI and opinion mining. IEEE Intelligent Systems, 25(3), 74-80, 2010. https://doi.org/10.1109/MIS.2010.75

[Choi, 00] Choi, S.-W.: Some statistical properties and Zipf's law in Korean text corpus. Journal of Quantitative Linguistics, 7(1), 19-30, 2000. https://doi.org/10.1076/0929-6174(200004)07:01;1-3;FT019

[Chung, 21] Chung, S.-H.: Applications of smart technologies in logistics and transport: A review. Transportation Research Part E: Logistics and Transportation Review, 153(102455), 2021. https://doi.org/10.1016/j.tre.2021.102455

[Corazza, 20] Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S.: A multilingual evaluation for online hate speech detection. ACM Transactions on Internet Technology, 20(2), 1-22, 2020. https://doi.org/10.1145/3377323

[Daugherty, 18] Daugherty, P. R., & Wilson, H. J.: Human+ machine: Reimagining work in the age of AI. Harvard Business Press, 2018.

[Dinah, 97] Dinah, P., Cecily, R., & Jorn, A.: A global code of Business Ethics. Journal of Business Ethics, 16(16), 1727-1735, 1997. https://doi.org/10.1023/A:1005877726316

[Farisco, 20] Farisco, M., Evers, K., & Salles, A.: Towards establishing criteria for the ethical analysis of artificial intelligence. Science and Engineering Ethics, 26, 2413-2425, 2020. https://doi.org/10.1007/s11948-020-00238-w

[Felzmann, 19] Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A.: Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 6(1), 2019. https://doi.org/10.1177/2053951719860542

[Fromm, 18] Fromm, J., Melzer, S., Ross, B., & Stieglitz, S.: Trump versus Clinton: Twitter communication during the US primaries. Network Intelligence Meets User Centered Social Media Networks 4, 201-217, 2018.

[Goodfellow, 17] Goodfellow, I.: Deep learning. MIT Press, Cambridge, MA, 2017. http://snu-primo.hosted.exlibrisgroup.com/82SNU:82SNU_INST21558843250002591

[Huang, 22] Huang, C., Zhang, Z., Mao, B., & Yao, X.: An overview of artificial intelligence ethics. IEEE Transactions on Artificial Intelligence, 1(1), 1-21, 2022. https://doi.org/10.1109/TAI.2022.3194503

[Jang, 19] Jang, B., Kim, I., & Kim, J. W.: Word2vec convolutional neural networks for classification of news articles and tweets. PloS one, 14(8), e0220976, 2019. https://doi.org/10.1371/journal.pone.0220976

[Johnson, 23] Johnson, D. G., & Verdicchio, M.: Ethical AI is Not about AI. Communications of the ACM, 66(2), 32-34, 2023. https://doi.org/10.1145/3576932

[Kane, 14] Kane, G. C., Alavi, M., Labianca, G., & Borgatti, S. P.: What's different about social media networks? A framework and research agenda. MIS Quarterly, 38(1), 275-304, 2014. https://doi.org/10.25300/MISQ/2014/38.1.13

[Kim, 22] Kim, N.-Y.: Suicide deaths of volleyball player, YouTuber rekindle calls for tougher punishment against cyberbullying, 2022. https://en.yna.co.kr/view/AEN20220208006300315

[Kuznetsov, 19] Kuznetsov, V., & Mariet, Z.: Foundations of Sequence-to-Sequence Modeling for Time Series. arXiv:1805.03714, 2019. https://doi.org/10.48550/arXiv.1805.03714

[Kyriacou, 16] Kyriacou, C., & Zuin, A.: Cyberbullying of teachers by students on YouTube: challenging the image of teacher authority in the digital age. Research Papers in Education, 31(3), 255-273, 2016.

[López-Meneses, 20] López-Meneses, E., Vázquez-Cano, E., González-Zamar, M.-D., & Abad-Segura, E.: Socioeconomic effects in cyberbullying: Global research trends in the educational context. International Journal of Environmental Research and Public Health, 17(12), 1-29, 2020. https://doi.org/10.3390/ijerph17124369

[Luo, 19] Luo, X., Tong, S., Fang, Z., & Qu, Z.: Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. Marketing Science, 38(6), 937-947, 2019.

[Makridakis, 17] Makridakis, S.: The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures, 90, 46-60, 2017.

[Malmasi, 18] Malmasi, S., & Zampieri, M.: Challenges in discriminating profanity from hate speech. Journal of Experimental & Theoretical Artificial Intelligence, 30(2), 187-202, 2018.

[Matalon, 21] Matalon, Y., Magdaci, O., Almozlino, A., & Yamin, D.: Using sentiment analysis to predict opinion inversion in Tweets of political communication. Scientific Reports, 11(1), 1-9, 2021. https://doi.org/10.1038/s41598-021-86510-w

[McCorduck, 04] McCorduck, P., & Cfe, C.: Machines who think: A personal inquiry into the history and prospects of artificial intelligence. CRC Press, 2004.

[Modha, 20] Modha, S., Majumder, P., Mandl, T., & Mandalia, C.: Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. Expert Systems with Applications, 161, 113725, 2020. https://doi.org/10.1016/j.eswa.2020.113725

[Muller, 21] Muller, H., Mayrhofer, M. T., Van Veen, E.-B., & Holzinger, A.: The ten commandments of ethical medical AI. Computer, 54(07), 119-123, 2021. https://doi.org/10.1109/MC.2021.3074263

[Munir, 18] Munir, M., Siddiqui, S. A., Dengel, A., & Ahmed, S.: DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. IEEE Access, 7, 1991-2005, 2018. https://doi.org/10.1109/ACCESS.2018.2886457

[Nam, 19] Nam, J.-E.: Cyberbullying and serial celebrity suicides. Hankoreh, 2019. https://english.hani.co.kr/arti/english_edition/e_national/918572.html

[Peebles, 14] Peebles, E.: Cyberbullying: Hiding behind the screen. Paediatrics & Child Health, 19(10), 527-528, 2014. https://doi.org/10.1093/pch/19.10.527

[Perols, 11] Perols, J.: Financial statement fraud detection: An analysis of statistical and machine learning algorithms. Auditing: A Journal of Practice & Theory, 30(2), 19-50, 2011. https://doi.org/10.2308/ajpt-50009

[Ribeiro, 20] Ribeiro, A. H., Tiels, K., Aguirre, L. A., & Schön, T.: Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness. International Conference on Artificial Intelligence and Statistics, 2020.

[Siau, 20] Siau, K., & Wang, W.: Artificial intelligence (AI) ethics: ethics of AI and ethical AI. Journal of Database Management, 31(2), 74-87, 2020. https://doi.org/10.4018/JDM.2020040105

[Singer, 09] Singer, J. B., & Ashman, I.: "Comment is free, but facts are sacred": User-generated content and ethical constructs at the Guardian. Journal of Mass Media Ethics, 24(1), 3-21, 2009. https://doi.org/10.1080/08900520802644345

[Song, 21] Song, T.-M., & Song, J.: Prediction of risk factors of cyberbullying-related words in Korea: Application of data mining using social big data. Telematics and Informatics, 58, 101524, 2021. https://doi.org/10.1016/j.tele.2020.101524

[Sun, 19] Sun, X., Gao, Y., Sutcliffe, R., Guo, S.-X., Wang, X., & Feng, J.: Word representation learning based on bidirectional grus with drop loss for sentiment classification. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(7), 4532-4542, 2019. https://doi.org/10.1109/TSMC.2019.2940097

[Sutskever, 14] Sutskever, I., Vinyals, O., & Le, Q. V.: Sequence to Sequence Learning with Neural Networks. arXiv preprint arXiv:1409.3215, 2014. https://doi.org/10.48550/arXiv.1409.3215

[Van Huynh, 19] Van Huynh, T., Nguyen, V. D., Van Nguyen, K., Nguyen, N. L.-T., & Nguyen, A. G.-T.: Hate speech detection on Vietnamese social media text using the bi-gru-lstm-cnn model. arXiv preprint arXiv:1911.03644, 2019. https://doi.org/10.48550/arXiv.1911.03644

[Wang, 21] Wang, P., Fan, E., & Wang, P.: Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recognition Letters, 141, 61-67, 2021. https://doi.org/10.1016/j.patrec.2020.07.042

[Wang, 14] Wang, S., Zheng, Z., Wu, Z., Lyu, M. R., & Yang, F.: Reputation measurement and malicious feedback rating prevention in web service recommendation systems. IEEE Transactions on Services Computing, 8(5), 755-767, 2014. https://doi.org/10.1109/TSC.2014.2320262

[Wang, 19] Wang, W., & Siau, K.: Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. Journal of Database Management, 30(1), 61-79, 2019. https://doi.org/10.4018/JDM.2019010104

[Wang, 20] Wang, Z., Zhou, J., Ma, J., Li, J., Ai, J., & Yang, Y.: Discovering attractive segments in the user-generated video streams. Information Processing & Management, 57(1), 102130, 2020. https://doi.org/10.1016/j.ipm.2019.102130

[Wattal, 10] Wattal, S., Schuff, D., Mandviwalla, M., & Williams, C. B.: Web 2.0 and politics: the 2008 US presidential election and an e-politics research agenda. MIS Quarterly, 669-688, 2010. https://doi.org/10.2307/25750700

[Yin, 17] Yin, W., Kann, K., Yu, M., & Schütze, H.: Comparative study of CNN and RNN for natural language processing. arXiv preprint arXiv:1702.01923, 1-7, 2017. https://doi.org/10.48550/arXiv.1702.01923

[Zhang, 21] Zhang, Q., Lu, J., & Jin, Y.: Artificial intelligence in recommender systems. Complex & Intelligent Systems, 7, 439-457, 2021. https://doi.org/10.1007/s40747-020-00212-w

[Zou, 18] Zou, J., & Schiebinger, L.: AI can be sexist and racist—it's time to make it fair. Nature, 559(7714), 324-326, 2018.

[Zulqarnain, 19] Zulqarnain, M., Ghazali, R., Ghouse, M. G., & Mushtaq, M. F.: Efficient processing of GRU based on word embedding for text classification. International Journal on Informatics Visualization, 3(4), 377-383, 2019. https://doi.org/10.30630/joiv.3.4.289