[1] Jose Alfredo de Vera III, M.S*

[2] John Paul Vergara, Ph. D.

The Impact of Russian Troll Tweets: Analyzing Political Motivation in Tweets from the Internet Research Agency

Journal of Electrical Systems

Abstract: - This research investigates the impact of social media trolling on political events such as elections. The study utilizes the dataset of the Internet Research Agency, a Russian "troll factory" indicted by the US Justice Department in February 2018, to analyze tweets from 2012 to 2018, with the aim of creating a classification algorithm that will predict the political motivation of future tweets. The study involves running various classification algorithms, including Naive Bayes, K Nearest Neighbors, Random Forest, Histogram-based Gradient Boosting Classification Tree, and Light Gradient Boosting Machine. Data cleaning and categorization were done, assigning numeric values to each category. The resulting model scores and a breakdown of precision per category were analyzed. Word frequency analysis was also conducted to identify the most frequent words and understand the overall sentiment of the tweets and the topics targeted by the trolls. The research findings indicate that predicting the category of tweets is challenging, but the classification algorithms were able to do so with relatively good accuracy, especially for the NewsFeed and HashtagGamer categories. The study emphasizes the need for further research to improve accuracy of tweet classification algorithms and for social media platforms to continue their efforts to identify and remove troll accounts that aim to undermine political bias.

Keywords: classification algorithms, machine learning, social media, troll farms, tweets.

I. INTRODUCTION

The Internet has been a tool that was initially conceptualized to bridge the geographical limitations of interconnectivity (Castells, 1997; Dutton, 1999; Greenstein & Downes, 1999). From its beginnings as research links between universities, it has evolved into a world with an overabundance of information, with various, and sometimes unfounded, sources of information. Social media has been the main channel of information dissemination and exchange, whether the information is a fact or opinion.

With his digital frontier comes the internet "troll", a user that posts content that is mainly outrageous, offensive, and opinionated, to stimulate a high arousal of emotion to their readers (Hardaker, 2013; Toker & Caner, 2016; Buckels, Trapnell, & Paulhus, 2014). Most of the time this is done simply to gain more engagement, which is commonly one of the metrics of social media topics, or offensive messages in online discussion forums or chat rooms with the intention of disrupting the conversation and eliciting emotional responses from other users.

The rise of trolls in online spaces has been linked to the affordances of technology and the way that these spaces are designed and moderated. For example, the anonymity afforded by many online platforms can make it easier for trolls to engage in harassing or disruptive behavior (Papacharissi, 2010). Additionally, the lack of clear guidelines and enforcement mechanisms on some platforms can contribute to a culture of permissiveness towards trolling behavior (Gillespie, 2018).

Trolls often target vulnerable groups, such as minorities or people with disabilities, in an attempt to elicit emotional responses from other users (Buckels, Trapnell, & Paulhus, 2014). This behavior can have serious consequences for those who are targeted, including psychological harm and increased feelings of social isolation (Wang, Chen, & Liang, 2011).

As the internet has reached more and more people, trolling has evolved to include various ways of provocation, including cyberbullying, doxing the creation of fake accounts, to anonymously harass and intimidate others. Despite efforts to regulate this trolling behaviour, it remains a significant problem on many social media platforms (Tandoc, Lim, & Ling, 2018). Some researchers have suggested that this is due in part to the way that social media metrics, such as likes and shares, incentivize the posting of provocative or controversial content (Marwick & Lewis, 2017).

During the 2016 United States presidential election, suspected Russian troll activity on elections has been a subject of significant concern due to its influence on the election outcome. There is evidence that Russian troll farms, which are organized groups of paid individuals created to disseminate false or misleading information online, interfered with the democratic processes not just in the United States, but also in other countries.

[1] Department of Information Systems and Computer Science, Ateneo de Manila University, Philippines. jdevera@ateneo.edu

[2] Department of Information Systems and Computer Science, Ateneo de Manila University, Philippines . jpvergara@ateneo.edu

* Corresponding Author Email: Philippines. jdevera@ateneo.edu

This disinformation was particularly effective in the election context, where emotions and opinions are already running high. By spreading false information, Russian trolls exacerbated existing political and social tensions, and created further division within a society. Russian trolls were also found to engage in other tactics to influence elections, such as spreading negative information about specific candidates or political parties or creating fake social media accounts to pretend to be ordinary citizens voicing their opinions. They sought to suppress voter turnout by spreading false information about the voting process, or by creating confusion or chaos around election day.

The impact of Russian troll activity on the 2016 United States presidential election was difficult to measure, as it is often difficult to discern the exact impact of online disinformation on individual voters. However, anecdotal reports emphasize how Russian troll activity undermined the democratic process by creating a climate of fear, uncertainty, and doubt. As such, many countries are now taking steps to combat Russian troll activity and other forms of online disinformation, through a variety of means, including regulation, education, and technological solutions.

II. Literature Review

Social media has emerged as a powerful platform that revolutionizes the way individuals exercise their civic rights and engage in political discourse. In the digital age, social media platforms have become vital tools for mobilizing communities, spreading awareness, and fostering democratic participation. Platforms like Twitter and Facebook have provided a voice to marginalized communities, fostering a more diverse and inclusive political discourse. These online platforms allow for the creation of virtual communities, facilitating network-building and coordination of collective actions. Movements such as the Arab Spring and #BlackLivesMatter have harnessed social media's power to mobilize people, raise awareness about systemic issues, and hold institutions accountable. Social media has become an accessible and effective tool for political engagement, empowering individuals from all backgrounds to participate in shaping their societies.

However, the influence of social media on politics is not without challenges. The rapid spread of misinformation and the manipulation of public opinion are significant concerns. The viral nature of social media facilitates the dissemination of false information, which can have far-reaching consequences on public perceptions and electoral processes.

Scholars have collectively emphasized the significant role of social media during elections. Literature on the prevalence of social media highlight how online platforms can contribute to political polarization, serve as channels for the spread of disinformation, and enable state-sponsored propaganda to influence public opinion. Kubin and von Sikorski (2021) found that social media platforms often facilitate the formation of echo chambers and filter bubbles, leading to limited exposure to diverse viewpoints and increased interaction among like-minded individuals. This selective exposure reinforces existing political biases and hampers the exchange of diverse perspectives, thus intensifying political polarization within online communities. Ong and Cabañes (2017) looked into how organizational structures and strategies employed by troll accounts in spreading fabricated news. Social media's rapid and extensive reach allows the swift dissemination of such content, potentially influencing public opinion and contributing to the spread of misinformation. This phenomenon undermines trust in established institutions and exacerbates political polarization among citizens.

In a similar vein, Golovchenko, Buntain, Eady, Brown, and Tucker (2020) contribute to this discourse by examining the influence of state-sponsored propaganda on public opinion through online platforms. Their study focuses on the impact of Russian troll accounts during the 2016 U.S. presidential election. These actors strategically exploited divisive issues and employed manipulation techniques to enhance the visibility and engagement of their propaganda content on platforms like Twitter and YouTube. Consequently, state-sponsored actors can effectively disseminate their messages and shape public opinion, potentially intensifying political polarization within online communities.

Emerging literature supports the notion that social media platforms play a significant role during elections, with the potential to contribute to political polarization, serve as channels for the spread of disinformation, and enable state-sponsored propaganda to influence public opinion. While social media provides a platform for democratic participation and mobilization, addressing the challenges posed by disinformation and polarization is crucial for fostering an informed, inclusive, and resilient democratic environment.

2.1 Investigating the Influence of Social Media in Politics: Diverse Research Approaches

Investigating the influence of social media in politics is essential for safeguarding democratic principles and electoral processes, as it sheds light on how these platforms shape political engagement, voter behavior, and the spread of misinformation and disinformation. Understanding the dynamics between online behavior in social media platforms and political discourse allows societies to develop informed policies and strategies to foster an engaged and resilient citizenry in the digital age.

Within the last decade, scholars have extensively investigated the influence of social media in politics through diverse research methods and approaches. Content analysis has been employed to examine the type and tone of political content shared on social media platforms, as seen in studies like Ong and Cabañes' (2017) investigation of troll accounts and fake news production. Surveys and polls, such as those conducted by The Pew Research Center (2021, 2015), have gathered data on how social media usage affects political attitudes and behavior. Case studies, like the one by Golovchenko (2022), examines the impact of Ukraine's ban on Russian social media platforms. The study finds that the ban, implemented in response to concerns about the spread of propaganda and disinformation, had significant effects on social media usage and political discourse in the country. The ban led to a substantial reduction in the use of Russian social media platforms within Ukraine, resulting in decreased access to Russian-controlled information sources. Network analysis, employed in studies like Udanor et al. (2016) underscores the influence of social media on political campaigns and electoral processes in developing countries. Udanor's research finds that political actors leverage social media platforms to reach a broader audience, disseminate campaign messages, and mobilize supporters. Social media's viral nature and ability to target specific voter segments contribute to its effectiveness as a campaign tool. Sentiment analysis, as used in studies by Se Jung Park et al. (2011), examine the political profiles of users on the South Korean social networking site, Cyworld. Se Jung Park's study revealed that Cyworld served as a significant platform for political expression and engagement among its users. Many individuals use their profiles to discuss and share political opinions, fostering political discourse and awareness within the online community.

Classification method is another approach that is particularly valuable when analyzing large datasets from social media platforms to understand how social media shapes political discussions. By categorizing data from online platforms, this methodology offers insights into how social media impacts political communication, behavior, and public opinion. Sentiment analysis through classification allows for understanding the emotional tone of political discussions on social media and how it influences public perception (Jürgens & Schoen, 2012). Political affiliation classification helps identify how users align themselves politically and whether it affects the content they engage with and share. By classifying posts based on accuracy and reliability, researchers can analyze the spread of misinformation and propaganda during political events (De Souza, et al., 2020). Classification also helps identify patterns of political mobilization, prevalent political issues, and the level of polarization within social media discussions (Frasincar, et a., 2017). Furthermore, it allows researchers to identify influential users and assess their impact on shaping public opinion (Koksal & Akgul, 2022; Park, et al., 2011). The use of classification in investigating social media's influence on politics provides valuable quantitative insights from vast datasets, complementing qualitative analyses and contributing to a deeper understanding of the interplay between social media and political dynamics.

These rigorous investigations have provided valuable insights into how social media platforms impact political communication, attitudes, behaviors, and outcomes, shedding light on the multifaceted role of social media in shaping the political landscape. As social media continue to be an integral part of daily discourse, robust examination of their impact on politics is crucial to protect democratic values and electoral integrity. The various methodologies offer valuable insights into the ways social media influences political participation, voter conduct, and the dissemination of false information and propaganda. Understanding the intricate relationship between online interactions on social media platforms and political discussions empowers societies to formulate well-informed policies and approaches to nurture an enlightened, involved, and resilient users in the era of digital communication.

III. METHODOLOGY

This research utilizes data from the Internet Research Agency, a Russian "troll factory" indicted by the US Justice Department in February 2018, to predict the political motivation of future tweets. The dataset comprises approximately 3 million tweets sent from Twitter handles associated with the Internet Research Agency, spanning the period from February 2012 to May 2018, with a significant concentration of tweets between 2015 and 2017.

The initial step involved data cleaning, which includes the removal of irrelevant features, such as tco*_step1 and id. The data is then preprocessed to ensure suitability for analysis. One hot encoding is employed to convert text in the categories field into numerical data, allowing for statistical analysis. Each unique category is assigned a numeric equivalent through this process.

Classification algorithms are employed to train the model that predicts the category of troll tweets. Multiple classification algorithms, namely Naive Bayes, K Nearest Neighbors, Random Forest, Histogram-based Gradient Boosting Classification Tree, and Light Gradient Boosting Machine, are utilized to train a predictive model that categorizes troll tweets. Model performance is evaluated through 5-fold cross validation to assess accuracy and consistency.

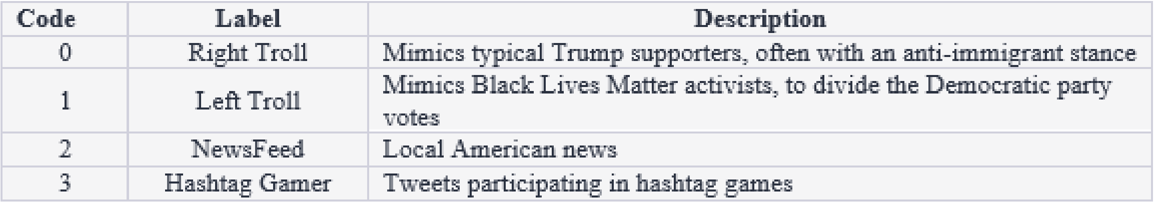
Following model evaluation, the resulting scores are analyzed, and a breakdown per category is examined.

Word frequency analysis is conducted to identify the occurrence of each word in the dataset, shedding light on the most frequent words used in tweets. This analysis aids in understanding the overall sentiment and targeted topics employed by the trolls. The tweet content is further refined by removing links and special characters. Subsequently, lemmatization is applied to reduce each word to its root form. Additionally, all stop words are removed from the corpus to enhance the analysis process.

IV. Research Findings and Discussion

One hot encoding was utilized to convert text in the categories field into numerical data for statistical analysis. The classification labels were converted into a numeric equivalent to enable proper classification by the algorithm. In the study, the labels were arbitrarily assigned the values 0 - 3 as shown in Table I.

Table I. Classification of Troll Tweets

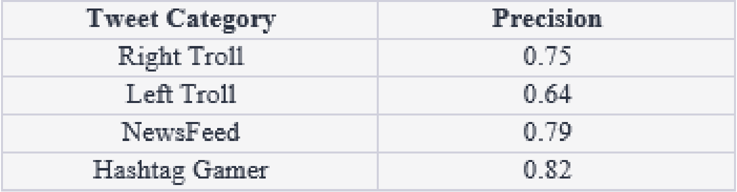


Code	Label	Description
0	Right Troll	Mimics typical Trump supporters, often with an anti-immigrant stance
1	Left Troll	Mimics Black Lives Matter activists, to divide the Democratic party votes
2	NewsFeed	Local American news
3	Hashtag Gamer	Tweets participating in hashtag games

The "Right Troll" category comprises accounts that mimic typical Trump supporters and frequently adopt an anti-immigrant stance. These trolls strategically aim to influence public opinion on right-leaning political and ideological issues, exacerbating divisions and polarizing discussions within the political landscape. On the other hand, the "Left Troll" accounts imitate Black Lives Matter activists to divide Democratic party votes. By assuming the appearance of genuine activists, they seek to exploit existing political rifts and foster internal discord within the Democratic party, thereby undermining progressive causes. In contrast, the "NewsFeed" troll accounts pose as sources of local American news. Unlike other troll categories, their content appears relatively neutral and informative, but the research suggests that they may still be involved in disseminating selective information or biased narratives, albeit in a subtler manner. The "Hashtag Gamer" trolls engage in participating in hashtag games on Twitter, using popular hashtags to insert disinformation or misleading content into broader conversations. By leveraging trending topics, these trolls aim to gain visibility and propagate their narratives through hashtag-driven discussions.

After applying different classification algorithms, a discernible pattern emerges in the precision of predicting the four major tweet categories as shown in Table II.

Table II. Model Precision Scores across the Four Tweet Classifications

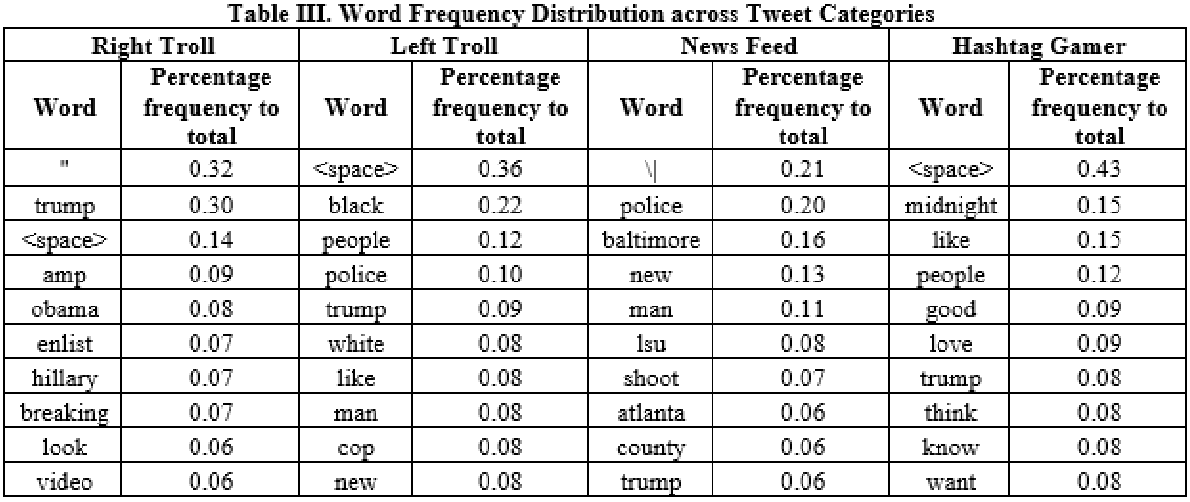


Tweet Category	Precision
Right Troll	0.75
Left Troll	0.64
NewsFeed	0.79
Hashtag Gamer	0.82

The model achieves an overall accuracy of 72% in classifying all the tweets. Precision for "Right Troll" signifies a 75% likelihood that a right troll tweet will be correctly identified. Similarly, the model demonstrates 64% precision in classifying "Left Troll" tweets, indicating its ability to correctly classify these tweets. For tweets categorized as "NewsFeed," the model performs with an precision of 79%, which is relatively high in this category. Lastly, the model excels in predicting "Hashtag Gamer" tweets, achieving the highest score of 82% precision. Along with the recall and F1 scores, these findings underscore the model's efficacy in effectively categorizing tweets, providing valuable insights into the distinct nature of Twitter content and user behavior across the identified categories. Among all the precision scores, it is observed that Left Troll tweets are the most ambiguous types of tweets, as it is the lowest among the other categories.

After conducting model evaluation, the model scores were analyzed, and the findings are presented in Table III, which showcases the Word frequency analysis and the occurrence of each word in the dataset across four tweet categories: Right Troll, Left Troll, News Feed, and Hashtag Gamer.

Table III. Word Frequency Distribution across Tweet Categories



Right Troll		Left Troll		News Feed		Hashtag Gamer	
Word	Percentage frequency to total	Word	Percentage frequency to total	Word	Percentage frequency to total	Word	Percentage frequency to total
"	0.32	<space>	0.36	\\|	0.21	<space>	0.43
trump	0.30	black	0.22	police	0.20	midnight	0.15
<space>	0.14	people	0.12	baltimore	0.16	like	0.15
amp	0.09	police	0.10	new	0.13	people	0.12
obama	0.08	trump	0.09	man	0.11	good	0.09
enlist	0.07	white	0.08	lsu	0.08	love	0.09
hillary	0.07	like	0.08	shoot	0.07	trump	0.08
breaking	0.07	man	0.08	atlanta	0.06	think	0.08
look	0.06	cop	0.08	county	0.06	know	0.08
video	0.06	new	0.08	trump	0.06	want	0.08

The table highlights specific words that appear uniquely in each tweet category, allowing for a deeper understanding of their distinctive characteristics. For instance, the word "obama" exclusively appears in Right Troll tweets, while "cop" is specific to Left Troll tweets. Additionally, the word "obama" constitutes 8% of all words in the Right Troll tweets, indicating its prevalence within that category.

Table III complements the precision scores in Table II by explaining the precision of the model algorithm. Notably, the Left Troll tweet category exhibits only three unique words specific to that group. The frequency percentage of these unique words is relatively low, with "black" appearing at 22%, and "white" and "cop" at 8%.

The table's comprehensive breakdown of word frequencies contributes to the model's effectiveness in accurately classifying tweets, enabling researchers to gain a more nuanced understanding of the distinct language patterns and themes associated with each tweet category. This understanding is instrumental in combating misinformation and disinformation, enhancing social media platform moderation, and promoting a more informed and responsible online discourse.

V. Conclusion

During the 2016 United States presidential election, the presence of suspected Russian troll activity became a significant concern due to its potential influence on the election outcome. These organized groups, known as Russian troll farms, were involved in disseminating false or misleading information online, not only within the United States but also in other countries. Their disinformation tactics were particularly effective in the emotionally charged election context, exacerbating existing political and social tensions and creating further division within society.

Assessing the exact impact of Russian troll activity on individual voters during the 2016 election proved challenging, as it is difficult to quantify the precise effects of online disinformation on voting behavior. Nevertheless, anecdotal reports highlighted how the presence of Russian trolls contributed to a climate of fear, uncertainty, and doubt, undermining the democratic process.

This research makes a significant contribution to the ongoing discussion on combatting troll activity and online disinformation through the innovative use of one hot encoding and classification algorithms. By efficiently converting text categories into numerical data, the research facilitates a deeper understanding of troll tweet patterns and their distinct characteristics. The model's precision scores in classifying different tweet categories, such as "Right Troll," "Left Troll," "NewsFeed," and "Hashtag Gamer," offer valuable insights into the prevalence and distribution of troll activity on social media platforms like Twitter. These findings play a crucial role in the development of more effective strategies to combat troll activity and online disinformation. By identifying and classifying troll tweets with precision, social media platforms and authorities can proactively target and remove harmful content, reducing the spread of disinformation and its potential impact on public discourse. The ability to accurately differentiate between different tweet categories also aids in the identification of potential troll accounts and the implementation of targeted measures to counter their activities.

These research findings have broader implications for combating online disinformation through various means. The study emphasizes the importance of regulatory efforts to hold social media platforms accountable for their content moderation practices. As we move forward in addressing these challenges, further studies should explore the potential of more advanced classification techniques and data processing methods to enhance our understanding of troll tweet patterns and user behavior.

References

[1] Ahmed H. Aliwy and Esraa H. Abdul Ameer (2017). Comparative Study of Five Text Classification Algorithms with their Improvements. Research India Publications. 4309-4319. http://www.ripublication.com

[2] Anderson, M., Rainie, L., Nolan, H., 2021. Social Media Use in 2021. Pew Research Center.

[3] Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. Personality and Individual Differences, 67, 97-102.

[4] Castells, M. (1997). The emergence of the internet and its implications for democratic society. In E. Katz & R. Rice (Eds.), Social consequences of internet use: Access, involvement, and interaction (pp. 5-24). MIT Press.

[5] De Souza, J. V., Gomes Jr, J., Souza Filho, F. M. de, Oliveira Julio, A. M. de, & de Souza, J. F. (2020). A systematic mapping on automatic classification of fake news in social media. Social Network Analysis and Mining, 10(1). https://doi.org/10.1007/s13278-020-00659-2

[6] Dutton, W. H. (1999). The social construction of reality in the information age. In E. N. Hahonou, D. C. Sharma, & Y. Waksman (Eds.), Information society: New media, ethics and postmodernism (pp. 7-25). Sage Publications.

[7] Frasincar, F., Ittoo, A., Nguyen, L. M., & Métais, E. (2017). Identifying Right-Wing Extremism in German Twitter Profiles: A Classification Approach. In Natural Language Processing and Information Systems (Vol. 10260, pp. 320–325). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-59569-6_40

[8] Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

[9] Golovchenko, Y. (2022). Fighting Propaganda with Censorship: A Study of the Ukrainian Ban on Russian Social Media. The Journal of Politics, 84(2), 639–654. https://doi.org/10.1086/716949

[10] Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A., & Tucker, J. A. (2020). Cross-Platform State Propaganda: Russian Trolls on Twitter and YouTube during the 2016 U.S. Presidential Election. The International Journal of Press/politics, 25(3), 357–389. https://doi.org/10.1177/1940161220912682

[11] Greenstein, S., & Downes, T. (1999). The evolution of the internet: From military experiment to general purpose technology. Journal of Policy Analysis and Management, 18(2), 322-343.

[12] Hardaker, C. (2013). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. Journal of Politeness Research, 9(1), 57-82.

[13] Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment" Social Science Computer Review, 30(2), 229–234. https://doi.org/10.1177/0894439311404119

[14] Kubin, E., & von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. Annals of the International Communication Association, 45(3), 188–206. https://doi.org/10.1080/23808985.2021.1976070

[15] Koksal, O., & Akgul, O. (2022). A Comparative Text Classification Study with Deep Learning-Based Algorithms. 2022 9th International Conference on Electrical and Electronics Engineering (ICEEE), 387–391. https://doi.org/10.1109/ICEEE55327.2022.9772587

[16] Linvill, D. L., & Warren, P. L. (2020). Troll Factories: Manufacturing Specialized Disinformation on Twitter. Political Communication, 37(4), 447–467. https://doi.org/10.1080/10584609.2020.1718257

[17] Marwick, A. E., & Lewis, R. (2017). Media manipulation and disinformation online. Data & Society Research Institute.

[18] Ong, J. C., & Cabañes, J. V. A. (2017). Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines. Media International Australia, 165(1), 71-87.

[19] Papacharissi, Z. (2010). A networked self: Identity, community, and culture on social network sites. Routledge.

[20] Perrin, A. (2015). "Social Networking Usage: 2005-2015." Pew Research Center. October 2015. Available at: http://www.pewinternet.org/2015/10/08/2015/Social-Networking-Usage-2005-2015/

[21] Se Jung Park, Yon Soo Lim, Sams, S., Sang Me Nam, & Han Woo Park. (2011). Networked Politics on Cyworld: The Text and Sentiment of Korean Political Profiles. Social Science Computer Review, 29(3), 288–299. https://doi.org/10.1177/0894439310382509

[22] Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news." Digital Journalism, 6(2), 137-153.

[23] Toker, Z., & Caner, A. (2016). Internet trolling and its impact on the workplace. Journal of Business Research, 69(10), 4344-4351.

[24] Udanor, C., Aneke, S., & Ogbuokiri, B. O. (2016). Determining social media impact on the politics of developing countries using social network analytics. Program : Electronic Library and Information Systems, 50(4), 481–507. https://doi.org/10.1108/PROG-02-2016-0011

[25] Wang, J., Chen, W., & Liang, Y. (2011). The effects of social media on college students. Journal of Educational Technology Development and Exchange, 4(1), 1-14.