**Journal of Digital Literacy and Volunteering**
Vol. 2, No. 2, July 2024, pp. 61-66
EISSN: 2985-8518
URL : http://jurnal.relawantik.or.id/index.php/ict

# Research on Online Hate Speech Detection from Popper and Kuhn's Philosophical Perspective

**Rinda Cahyana [1*], Leni Fitriani [2], Yudi Setiawan [3], Dimitri Mahayana [4]**

[1,2] Departement of Informatics, Faculty of Computer Science, Institut Teknologi Garut
*rindacahyana@itg.ac.id, leni.fitriani@itg.ac.id*

[3] Department of Information System, Faculty of Engineering, Universitas Bengkulu
*ysetiawan@unib.ac.id*

[4] School of Electrical Enginering and Informatics, Institut Teknologi Bandung
*dimitri@office.itb.ac.id*

***Keywords:***

*Artificial Inteligence,*
*Computer Science,*
*Hate Speech,*
*Philosophy,*
*Social Media*

***ABSTRACT***

*The negative impact of spreading hate speech on social media has prompted various parties to intervene. Computer science researchers have conducted experiments to find solutions for automated intervention by applying artificial intelligence, such as machine learning and deep learning. The fulfillment of the theory of truth makes the machine learning paradigm considered by scientists to solve problems. However, the increasing size of social media data has shifted its paradigm to deep learning. Deep learning becomes a new normal science after completing the task of classifying hate speech well on a large amount of data. However, any approach will be an anomaly when it cannot complete the task. The accessibility of research resources makes it easier for researchers to determine the nature of their experiments, whether scientific or pseudo-science.*

**Corresponding Author :**

Rinda Cahyana,
Institut Teknologi Garut,
Mayor Syamsu street 1, Garut, Indonesia, 44151
Email: rindacahyana@itg.ac.id

## 1. INTRODUCTION

Social media applications allow anyone to communicate anywhere [1]. Communication in the interpersonal context involves several people producing and processing messages [2]. Messages in text, image, and sound formats are spread across social media [3] and flow instantly between senders and recipients [4]. Among these messages are cyberhate [5], [6], [7] and cyberbullying [8].

Hate speech is an offensive communication mechanism that disparages a person or group based on protected innate characteristics, such as race, color, ethnicity, gender, disability, sexual orientation, nationality, religion, political affiliation, and so forth [9], [10], [11]. Bullying is an adverse action carried out by individuals or groups directly and repeatedly [12]. Cyber words that complement hate and bullying indicate cyberspace or the internet as a place for spreading the message.

Cyberbullying is rife in the world of education [13], [14] as a result of the lack of a teacher or school administrator's role in identifying its existence [15]. Cyberbullying has terrible effects, such as physical and mental health problems and death [16]. In contrast to cyberbullying, which impacts a specific target, cyberhate has a broad impact [17]. The effect of cyberhate can develop into exclusivism, physical attacks, and extermination of outside groups [18].

Various intervention efforts to deal with hate speech content on social media include the enforcement of laws [19], [20], and counter-speech [21]. Victims can report on social media applications [22] after receiving the harmful content and feeling the bad effects. Another handling approach gives everyone the authority to quarantine the content to avoid its dangerous effects [23].

Currently, many studies use automatic hate speech detection as an intervention strategy [24]. However, artificial intelligence systems can make wrong conclusions [25], partly because of different definitions of hate speech or the diversity and limitations of data [26]. These limitations make the experimental paradigm of automatic hate speech detection continue to change and place the experiment between the positions of science and pseudo-science. This study aims to explain these changes and positions with the philosophical thought of Karl Popper and Thomas Kuhn. This research aims to explain changes in the paradigm of automatic hate speech detection and the position of experiments between science and pseudoscience with the philosophical thoughts of Karl Popper and Thomas Kuhn. The problem formulation is as follows:

1. What is Popper and Kuhn's philosophical thinking regarding paradigm shifts?
2. How do paradigm shifts occur in automatic hate speech detection research?
3. Under what conditions do automatic hate speech detection experiments fall under the conditions of science and Pseudoscience?

## 2. RESEARCH METHOD

The literature survey is part of the activities of *the scientific field review* stage in the *sequential research process* [27], [28]. The method is as shown in Figure 1. This research uses a literature survey to achieve the objectives. The definition search includes limitations of literature searches based on the topic of interest, its popularity, type of publication, and quality. The literature search pays attention to these definitions to obtain material for critical evaluation according to needs. Critical evaluation of selected material will provide an understanding of the subject area and written material in the literature review. Material evaluation produces a list of relevant bibliography to achieve research objectives.

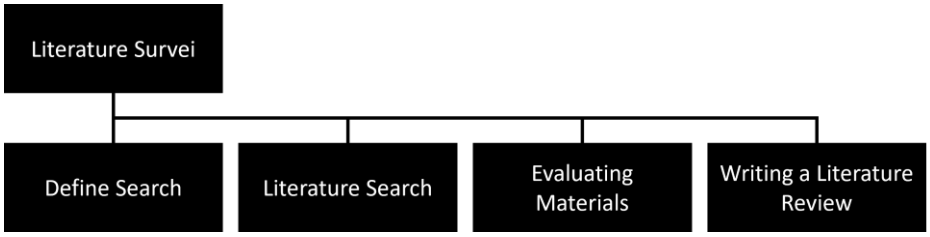


**Figure 1.** Research Stages

Hate speech topics cover areas of legal and social science knowledge and computer science and engineering [29]. In computer science, online hate speech detection system is a new research [17]. This area is the scope of the literature search in this research. The library sources used are scientific journal articles, proceedings articles, and books.

## 3. RESULTS AND ANALYSIS

This research selected 19 materials as references to achieve the research objectives, which included 7 books, 10 journal articles with an international reputation in the 1st quartile, and 1 proceedings article. Analysis of the evaluation results of materials is divided into three parts, according to the problem formulation.

### 3.1 Popper and Kuhn's Philosophical Thoughts

Karl Popper is one philosopher who rejects empirical science that uses inductive reasoning. Popper argues that particular statements cannot be universal, so conclusions from specific experiences do not represent general experiences. He proposed a scientific test method based on deduction and falsification, in

which a hypothesis P becomes false if there are implications (p1, p2, ... pn) that are proven false [30]. As long as the experiment does not prove that p is wrong, then P is temporarily corroborated.

Even though the deductive approach seems incompatible with the inductive, the two approaches complement each other. For example, deductive reasoning using the theory of gravity to predict satellite trajectories will not be solid if there is no contribution from Relativistic Physics and confirmation or repeated testing of Newton's general hypothesis using new examples [31]. Newton made a general hypothesis through inductive learning that specifically observed a sample of data in the form of a falling apple. An illustration of the integration of these two approaches is shown in Figure 2.

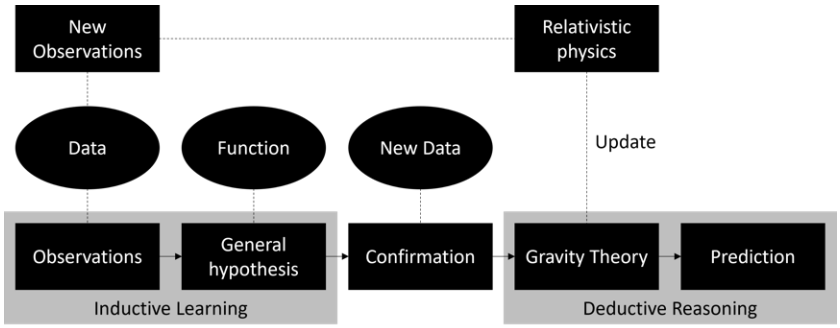


**Figure 2.** Newton's General Hypothesis Strengthening Process

A theory or claim becomes pseudo-science if it rejects the test or does not pass the test. Test rejection can occur due to the following fallacy: 1) *Argument from incredulity*, failing to understand or refusing to believe the test results; and 2) *Argumentum ad verecundiam*, the attitude of not wanting intellectuals who make claims appear defeated. Thomas Kuhn had the same thoughts as Popper about relative knowledge. Kuhn argues that there are always anomalies, namely scientific problems that cannot be completely solved by normal science [32]. This pile of anomalies will create a crisis that the new normal science will one day solve. The process of replacing new normal science will occur continuously, so humans will never be able to create absolute knowledge or paradigms.

### 3.2 The Shifting of Paradigm

Artificial intelligence is a research area or field of science that allows computers to imitate human qualities or do things that humans currently do well, such as learning, reasoning, communicating, seeing, and hearing [33], [34], [35], [36]. In the method of reasoning, artificial intelligence uses a knowledge base to make hypotheses, and the predictions it produces are highly interpretable by humans. In contrast, in the learning method, artificial intelligence uses a lot of data to make predictions that humans cannot interpret, so there is no need to study the data [31].

Machine learning is a form of artificial intelligence where the system can improve its task performance by studying data or previous task experience without following explicit instructions [37], [38]. The application of machine learning in hate speech detection on Twitter includes the stages of data collection, labeling, data splitting, extracting features, adding knowledge from external sources, machine learning, and measuring accuracy [39]. Twitter data can be collected by Spark using the Twitter-API and stored in big data platforms such as Hadoop so that the data is stored and replicated on multiple servers [40]. Twitter-API is Twitter's data access framework, an alternative data collection method besides Open Datasets and Social Honeypots [41].

From 2014 to 2017, the Support Vector Machine (SVM) was more famous than deep learning for hate speech detection [29]. SVM can be explicitly applied to detect hate speech that targets various protected characteristics, such as religion, race, disability, and sexual orientation [42]. SVM performance can be better than other methods with the proper feature extraction method, for example, character n-gram [43].

However, SVM performance scores can be under deep learning in specific datasets and methods, such as Deep Convolutional Neural Networks [44]. The fusion approach can improve deep learning performance [45]. The advantage of deep learning over traditional machine learning is mainly due to its ability to analyze big data and unsupervised learning [46].

While traditional machine learning continues to be an anomaly for not excelling in online hate speech classification tasks that use large datasets, deep learning methods will become the new paradigm or normal

science. However, traditional machine learning remains useful in other classification tasks that use small datasets. The classifier model will continue to be used in experiments as long as its performance can still compete with the deep learning classifier model.

Research that uses specific methods by considering the performance achievements of these methods in previous studies is normal science. The process of changing paradigms to the next new normal science occurs continuously following the journey of the experiment. A consensus was formed among the researchers when they used the same method to consider its reputation for building the best classifier models.

### 3.3 Science and Pseudoscience

Disagreements can occur within one nation or between nations; for example, Americans think the application of the law to freedom of speech in Britain a restriction on the expression of freedom [47]. The boundary between hate speech and freedom of expression is blurred, causing the definition of hate speech to not be universally accepted. Whereas a clear definition of hate speech can make it easier to annotate hate speech [26].

Under such conditions, the output of all stages becomes useless. If we ignore this disagreement, research will be considered a science only because it has been tested after passing the stage of measuring accuracy; otherwise, it would be pseudo-science. Therefore, we cannot generalize the performance of classifiers using partially applicable data sets.

Research is a science as long as its research methodology includes practical-empirical testing. Meanwhile, valid research can become pseudo-science if it shows an attitude of not being willing to be tested by other research. The provision of open sources in scientific publications opens the opportunity for future research to repeat the experiment again and compare the results with the results of other experiments. Research resources that are not state secrets should be open so that other researchers can re-examine them and contribute by starting, continuing, or updating the results of previous research [48]. Artificial intelligence research related to the identification of negative content can fulfill the theory of truth as long as it fulfills the following conditions: 1) Correspondence, where research uses data, methods, and measurable performance calculations; 2) Coherence, where the researcher compares with other methods to make more objective conclusions; 3) Consensus, where research uses agreed methods that are suitable for the classification task it performs; and 4) Pragmatics, where the results of the research are useful for identifying the content of hate speech on social media.

## 4. CONCLUSION

This research has explained the paradigm shift according to the philosophical perspective of Thomas Kuhn and Popper, where new normal science succeeded in solving problems that could not be handled by previous normal science. In the context of automatic hate speech detection, a paradigm shift occurs, for example, when deep learning can handle classification tasks on big data better than traditional machine learning. Automatic hate speech detection research becomes pseudoscience if the researcher does not allow other researchers to test the results, for example, by closing access to related resources.

## ACKNOWLEDGMENTS

We want to thank the students in the philosophy course for providing helpful insights into this research, even though they may disagree with all of the interpretations or conclusions of this paper.

## REFERENCES

[1] W. N. H. W. Ali, M. Mohd, and F. Fauzi, "Cyberbullying detection: an overview," in *2018 Cyber Resilience Conference (CRC)*, IEEE, 2018, pp. 1–3.

[2] D. O. Braithwaite and P. Schrodt, *Engaging theories in interpersonal communication: Multiple perspectives*. Routledge, 2021.

[3] C. E. Ring, "Hate speech in social media: An exploration of the problem and its proposed solutions." University of Colorado at Boulder, 2013.
[4] M. S. Albarrak, M. Elnahass, S. Papagiannidis, and A. Salama, "The effect of twitter dissemination on cost of equity: A big data approach," *International Journal of Information Management*, vol. 50, pp. 1–16, 2020, doi: https://doi.org/10.1016/j.ijinfomgt.2019.04.014.
[5] J. Hawdon, A. Oksanen, and P. Räsänen, "Exposure to Online Hate in Four Nations: A Cross-National Consideration," *Deviant Behavior*, vol. 38, no. 3, pp. 254–266, Mar. 2017, doi: 10.1080/01639625.2016.1196985.
[6] S. Wachs and M. F. Wright, "The Moderation of Online Disinhibition and Sex on the Relationship Between Online Hate Victimization and Perpetration," *Cyberpsychology, Behavior, and Social Networking*, vol. 22, no. 5, pp. 300–306, Apr. 2019, doi: 10.1089/cyber.2018.0551.
[7] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *IEEE Access*, vol. 6, pp. 13825–13835, 2018.
[8] P. Sheldon, P. Rauschnabel, and J. M. Honeycutt, *The dark side of social media: Psychological, managerial, and societal perspectives*. Academic Press, 2019.
[9] N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," *Aggression and Violent Behavior*, vol. 40, pp. 108–118, 2018, doi: https://doi.org/10.1016/j.avb.2018.05.003.
[10] R. Cohen-Almagor, "Fighting Hate and Bigotry on the Internet," *Policy & Internet*, vol. 3, no. 3, pp. 1–26, Sep. 2011, doi: https://doi.org/10.2202/1944-2866.1059.
[11] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? the challenging case of long tail on twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, 2019.
[12] A. Bozyiğit, S. Utku, and E. Nasibov, "Cyberbullying detection: Utilizing social media features," *Expert Systems with Applications*, vol. 179, p. 115001, 2021.
[13] W. M. Al-Rahmi, N. Yahaya, M. M. Alamri, N. A. Aljarboa, Y. Bin Kamin, and M. S. Bin Saud, "How cyber stalking and cyber bullying affect students' open learning," *Ieee Access*, vol. 7, pp. 20199–20210, 2019.
[14] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 3–24, 2017.
[15] V. Banerjee, J. Telavane, P. R. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," *2019 5th International Conference on Advanced Computing \& Communication Systems (ICACCS)*, pp. 604–607, 2019.
[16] C. P. Barlett, "Chapter 2-Cyberbullying. Traditional bullying. and aggression: A complicated relationship," *Predicting Cyberbullying*, pp. 11–16, 2019.
[17] A. Al-Hassan and H. Al-Dossari, "Detection Of Hate Speech In Social Networks: A Survey On Multilingual Corpus," *Computer Science \& Information Technology(CS \& IT)*, 2019.
[18] M. Kopytowska and F. Baider, "From stereotypes and prejudice to verbal and physical violence: Hate speech in context," *Lodz Papers in Pragmatics*, vol. 13, no. 2, pp. 133–152, 2017, doi: doi:10.1515/lpp-2017-0008.
[19] A. Brown, "The Racial and Religious Hatred Act 2006: a Millian response," *Critical Review of International Social and Political Philosophy*, vol. 11, no. 1, pp. 1–24, Mar. 2008, doi: 10.1080/13698230701880471.
[20] C. Ezeibe, "Hate Speech and Election Violence in Nigeria," *Journal of Asian and African Studies*, vol. 56, no. 4, pp. 919–935, Sep. 2020, doi: 10.1177/0021909620951208.
[21] A. A. Siegel and V. Badaan, "#No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online," *American Political Science Review*, vol. 114, no. 3, pp. 837–855, 2020, doi: DOI: 10.1017/S0003055420000283.
[22] K. D. Sainju, N. Mishra, A. Kuffour, and L. Young, "Bullying discourse on Twitter: An examination of bully-related tweets using supervised machine learning," *Computers in human behavior*, vol. 120, p. 106735, 2021.
[23] S. Ullmann and M. Tomalin, "Quarantining online hate speech: technical and ethical perspectives," *Ethics and Information Technology*, vol. 22, no. 1, pp. 69–80, 2020, doi: 10.1007/s10676-019-09516-z.
[24] C. Blaya, "Cyberhate: A review and content analysis of intervention strategies," *Aggression and Violent Behavior*, vol. 45, pp. 163–172, 2019, doi: https://doi.org/10.1016/j.avb.2018.05.006.
[25] J. M. Helm *et al.*, "Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions," *Current reviews in musculoskeletal medicine*, vol. 13, no. 1, pp. 69–76, Feb. 2020, doi: 10.1007/s12178-020-09600-8.
[26] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PLOS ONE*, vol. 14, no. 8, p. e0221152, Aug. 2019.
[27] T. Greenfield, "Research methods: guidance for postgraduates.," *Research methods: guidance for postgraduates. London; Toronto: Arnold; Wiley*, 1996.
[28] J. A. Sharp, J. Peters, and K. Howard, *The management of a student research project*. Routledge, 2017.
[29] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," *ACM Comput. Surv.*, vol. 51, no. 4, 2018, doi: 10.1145/3232676.
[30] K. Popper, *The logic of scientific discovery*. Routledge, 2005.

[31] C. C. Aggarwal, *Artificial Intelligence: A Textbook*, 1st ed. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72357-6.
[32] D. Shapere, "The structure of scientific revolutions," *The Philosophical Review*, vol. 73, no. 3, pp. 383–394, 1964.
[33] S. Haag and P. Keen, *Information Technology: Tomorrow's Advantage Today.* McGraw-Hill Companies, Inc., 1996.
[34] P. W. Langley, H. A. Simon, G. Bradshaw, and J. M. Zytkow, "Scientific Discovery: Computational Explorations of the Creative Process." The MIT Press, Feb. 24, 1987. doi: 10.7551/mitpress/6090.001.0001.
[35] E. Rich, K. Knight, and S. B. Nair, *Artificial Intelligence*, 3rd ed. Tata McGraw-Hill Publishing Company, Ltd., 2009.
[36] B. K. Williams and S. C. Sawyer, *Using Information Technology: a Practical Introduction to Computer & Communications*. New York: McGraw-Hill, 2011.
[37] T. M. Mitchell, *Machine Learning*. McGraw-Hill Science, 1997.
[38] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM J. Res. Dev.*, vol. 3, no. 3, pp. 210–229, 1959, doi: 10.1147/rd.33.0210.
[39] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, "A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks," *Expert Systems with Applications*, vol. 164, p. 114006, 2021, doi: https://doi.org/10.1016/j.eswa.2020.114006.
[40] S. M. Alzahrani, "Big Data Analytics Tools: Twitter API and Spark," in *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, IEEE, 2021, pp. 1–6.
[41] N. S. Mullah and W. M. N. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," *IEEE Access*, vol. 9, pp. 88364–88376, 2021, doi: 10.1109/ACCESS.2021.3089515.
[42] P. Burnap and M. L. Williams, "Us and them: identifying cyber hate on Twitter across multiple protected characteristics," *EPJ Data Science*, vol. 5, no. 1, p. 11, 2016, doi: 10.1140/epjds/s13688-016-0072-6.
[43] O. Oriola and E. Kotzé, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," *IEEE Access*, vol. 8, pp. 21496–21509, 2020, doi: 10.1109/ACCESS.2020.2968173.
[44] P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao, "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," *IEEE Access*, vol. 8, pp. 204951–204962, 2020, doi: 10.1109/ACCESS.2020.3037073.
[45] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep Learning Based Fusion Approach for Hate Speech Detection," *IEEE Access*, vol. 8, pp. 128923–128929, 2020, doi: 10.1109/ACCESS.2020.3009244.
[46] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 1, 2015, doi: 10.1186/s40537-014-0007-7.
[47] J. W. Howard, "Free Speech and Hate Speech," *Annual Review of Political Science*, vol. 22, no. 1, pp. 93–109, May 2019, doi: 10.1146/annurev-polisci-051517-012343.
[48] C. W. Dawson, *Projects in Computing and Information Systems: A Student's Guide*. Pearson Prentice Hall, 2009.