The 7th International Conference on Economics and Social Sciences
Exploring Global Perspectives:
The Future of Economics and Social Sciences
June 13-14, 2024
Bucharest University of Economic Studies, Romania

Sentiment Analysis of Research on AI Ethics: A Web-Based Study

Alexandra-Cristina-Daniela CIUVERCA[1]

DOI: 10.24818/ICESS/2024/060

Abstract

The field of Artificial Intelligence has experienced significant growth in recent years, both in terms of technological development and global adoption rates. AI-based solutions are now finding their way into the lives of an increasing number of people worldwide, being used both professionally and personally. However, along with this progress, numerous controversies have arisen regarding the ethics of their use in various domains. In the specialised literature, there are a growing number of publications focused on the discussion of this topic. The present study focusses on analysing the general trends of these scientific works in relation to the ethics of the use of AI. Using the Web of Science Clarivate database, a set of publications was selected based on keywords and subsequently subjected to analysis. Sentiment analysis techniques are used to identify the positive or negative trend among specialists and how it has evolved over the years. Latent Dirichlet Allocation is used to highlight the main topics developed in these writings. On the basis of the obtained results, it can be noted that the concern for AI ethics issues is increasingly addressed in specialised writings. Sentiment analysis reveals that, in recent publications, on average, sentiments tend to be slightly positive, but the polarity value has decreased in recent years. Thus, this study contributes to a better understanding of specialists' positions regarding identified AI ethics issues, highlighting results obtained through the application of modern natural language processing techniques and by presenting important aspects emphasised in existing scientific works on this topic.

Keywords: artificial intelligence, ethics, research publication, latent Dirichlet allocation, sentiment analysis.

JEL Classification: C55; O33; I23.

[1] Bucharest University of Economic Studies, Bucharest, Romania, alexandra.ciuverca@csie.ase.ro.

1. Introduction

In recent years, marked by an indisputable development of Big Data and the Internet of Things (IoT), solutions based on artificial intelligence have begun to play an increasingly important role in both the lives of individuals and the activity of companies in all fields. Having the ability to be integrated in numerous and diverse scenarios, artificial intelligence has come to have a considerable impact on society, but this has also been accompanied by concerns regarding the ethics of using AI solutions (Etzioni & Etzioni, 2017). Previous studies in the specialised literature have analysed people's opinion expressed in the online environment regarding the ethics of AI. A study conducted in 2023 used data extracted from the YouTube and Twitter platforms to analyse the general trend of users' feelings regarding the ethics of using artificial intelligence. Following the application of methods based on the Naive Bayes Classifier and algorithms to calculate the frequency of terms and the inverse frequency of documents (TF-IDF), the study concluded that most of the texts analysed reflect a rather negative attitude, 62.4% of the scores obtained being negative (Yoga Saputra et al., 2023). An article published in 2022 subjected to analysis a set of data that included only user posts extracted from the Twitter social media platform, obtaining a result that indicates predominantly neutral feelings (Dwivedi et al., 2022). It is important to mention that many previous studies suggest the importance of addressing the problems and limitations related to AI ethics in order to increase the level of transparency and acceptance of AI solutions (Charan, 2023; Karoo & Chitte, 2023). The present study aims to investigate the opinion of the academic community regarding AI ethics by analysing abstracts of scientific indexed Web of Science works between 1991 and the year of publication of this work. For this purpose, multiple natural language processing (NLP) techniques were used. NLP methods represent a branch of AI that facilitates the dissemination and extraction of relevant information from large volumes of data with an appreciable level of precision. Among the forms of application of NLP algorithms are text processing, spelling checking, and error correction, methods of extracting and classification information from texts, answering questions, and automatic translation (Chowdhary, 2020). The modelling of the topics addressed was carried out by means of the Latent Dirichlet Allocation (LDA) method. In the following sections, the objectives of this study will be detailed, the application of the methods listed above will be described, and the results obtained will be presented and interpreted.

2. Problem Statement

As Gao et al. (2024) presented, in the absence of an ethically fitting trajectory for the development of artificial intelligence, the main associated risks are related to unpredictability, lack of transparency, damage to human privacy, and lack of fairness caused by bias. In the specialised articles, different particular vulnerabilities of the use of AI are highlighted depending on the field of applicability. In areas of activity such as national defence (Blanchard et al., 2024; Hadlington et al., 2024), cybersecurity (González et al., 2024), medicine (Heyen & Salloch, 2021), telemedicine (Pool et al., 2024), education (Leal Filho et al., 2024; Tossell et al., 2024), human resources (Hamilton & Davison, 2022), the importance of using artificial intelligence responsibly to maximise the level of efficiency and benefits is highlighted. For this purpose, the use of unbiased machine learning algorithms must be targeted, which maintain equity between people as individual beings, with specific needs, with their own vulnerabilities (Giovanola & Tiribelli, 2023). One such area that sparks many debates on the ethics of AI-based solutions is the field of medicine. Despite significant progress in AI solutions in recent years and many years since the application of machine learning algorithms in medicine began, patient safety, minimising the risk of errors or bias, and the efficiency of technologies remain at the forefront of concerns for specialists in all branches of medicine (Drabiak et al., 2023).

However, according to what was previously stated, the field of medicine is not the only field that has shown scepticism regarding the use of solutions based on artificial intelligence, and one of the reasons behind this scepticism is the ethical problems of AI. The academic community is increasingly addressing these problems through scientific work. Thus, through this study, it is desired to identify the main topics addressed by the specialists and the evolution of the community's feeling in relation to this topic. Therefore, the contributing elements of the current study that lead to the achievement of the previously stated goals are: (a) Analysis of the set of relevant publications for the chosen topic and identification of the related central topics using NLP techniques; (b) Sentiment analysis applied to scientific work, offering a new perspective and extending the conclusions previously obtained through research based on data from the social media environment (Yoga Saputra et al., 2023) or Wikipedia data (Wei et al., 2024); (c) Using the probabilistic LDA model, an advanced analytical technique of natural language processing, to identify the main topics. Subsequently, t-SNE was used to generate a graphical representation that allowed the visualisation and interpretation of the relationships and distances between the different subjects identified by LDA. Using t-SNE helps to convert complex topic distributions into a two-dimensional space. This makes the differences and similarities easy to see.

3. Research Questions / Objectives of the Research

This research aims to describe the trend in the specialised literature on the topic of AI ethics, highlighting how sentiments towards the chosen subject have evolved over time. Considering the results of sentiment analysis whose results were presented in the articles referred to in the introductory section of this work, it is desired to find out if the hypothesis will be accepted that the scientific community tends to express predominantly negative feelings towards neutral ones or if it has rather a positive attitude in relation to the ethics of using AI solutions. Also, it is desired that through the interpretation of the results obtained, it will be found out if the topics extracted really reflect concern and reluctance or if the topics addressed inspire confidence in the development and improvement of AI ethics. Given the hypotheses stated, this paper aims to answer three main questions:

(Q1) What are the most relevant and frequently discussed topics by the scientific community in relation to the ethics of artificial intelligence?

(Q2) What is the general trend of researchers' feelings regarding AI ethics?

(Q3) How have the feelings expressed by the authors of scientific works evolved over the years?

The contribution brought about by this paper lies in the approach to understanding the link established between the ethics of artificial intelligence and the perspective that the scientific community has regarding this aspect.

4. Research Methods

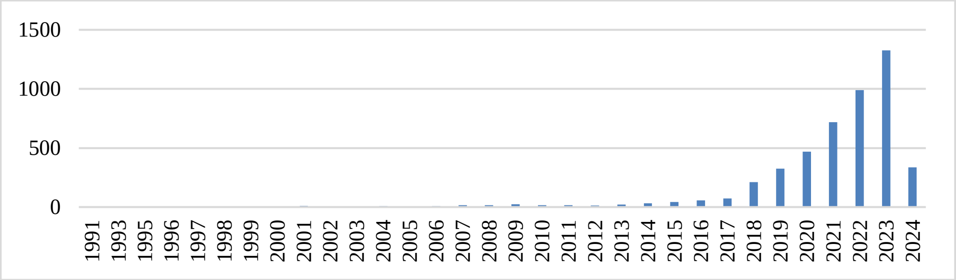
For analysis, a data set was utilised, exported from the Web of Science Clarivate platform, following the filtration of records through representative keywords for the topic of interest (“Artificial intelligence ethics,” “AI ethics,” “Ethics of AI”). This resulted in 5.041 records from the Web of Science Core Collection database, including the texts from the Abstract section of publications. 56% of the articles were published in European countries, 40% in the North American continent, and the remaining 4% had publishers from other continents. Among these, 279 were excluded from the analysis because they did not have an associated abstract. Therefore, a total of 4.762 abstracts were included in the analysis.

Each method applied in this process aims to extract, process, and visualise valuable information related to scientific articles related to AI ethics, through the use of abstracts, facilitating a better understanding of the content and feelings expressed. These techniques, from pre-processing to advanced modelling and visualisation, help transform data into knowledge. The first step consisted of the initial preparation for analysis of the texts in the Abstract section of each article by removing missing values and by pre-processing the text to remove noise (such as punctuation, common words, and stop words). Thus, the records containing null values in the “Abstract” column were eliminated. Then tokenization of texts, removal of stop words and punctuation, and lemmatisation of words were performed to reduce lexical variation. A WordCloud diagram was built as a visual representation of the frequency of words in the data set, in order to quickly identify the words and implicitly the key themes. The diagram was adjusted by excluding stop words to focus on keywords that add value to understanding the subject of the abstracts. Sentiment analysis was used to evaluate the general tone (positive, negative, neutral) of the abstracts using a pretrained sentiment analysis model. Using the TextBlob library to calculate the polarity scores for each individual abstract. Subsequently, the scores were aggregated to determine the prevailing sentiment for the entire data set. The evolution of feelings over the years was also tracked (the set includes articles published between 1991 and today). The main keywords used in the abstracts of scientific works were extracted by means of NLP techniques. Later, a HeatMap graph was built to highlight the prevalence of the different topics addressed over time. The main keywords used in the abstracts of scientific works were extracted by means of NLP techniques. Later, temporal analysis was performed, and a HeatMap-type graph was built to highlight the prevalence of the various themes addressed over time. Keyword frequency can outline the research trends of the scientific community and the changes that have occurred in recent years, in which the popularity of AI solutions has grown significantly. The topics were modelled with LDA (Latent Dirichlet Allocation). The aim is to identify and extract the main topics (topics) described in the abstracts. This helps us to understand the main themes and the informational structure of the data set. An LDA model was generated to extract groups of words that frequently appear together in documents and compose topics.

5. Findings

It is well-known that the interest in AI solutions has grown exponentially in recent years, together with the notable advances in the field. This is also reflected in the number of scientific articles published by the academic community, which is continuously increasing. When analysing the abstracts of the works indexed on the Web of Science, this upward trend is confirmed. As can be observed in Figure 1, located below, the number of publications on this topic has increased significantly over the years, along with the evolution of AI solutions and their adoption in multiple fields of activity.

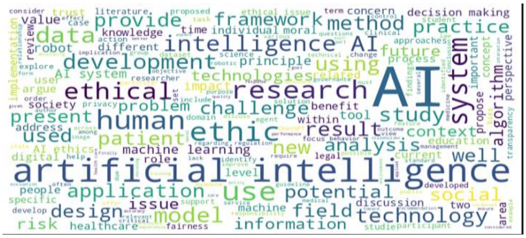
Figure 1. The number of publications in the specialised literature on topics related to artificial intelligence ethics published annually



Source: Web of Science https://www.webofscience.com/ (accessed April 2024).

To highlight keywords from the extracted data set and to understand the existing trends, a WordCloud time graph (as in Figure 2) was built using the Python library of the same name.

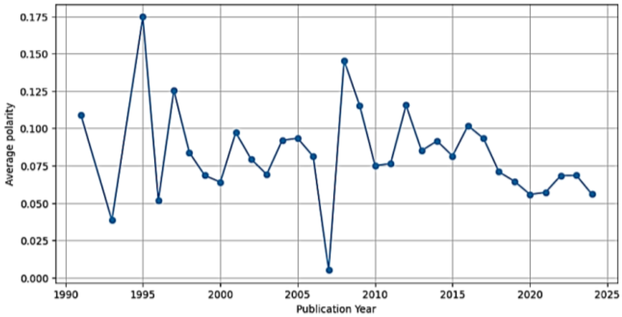
Figure 2. WordCloud applied to the abstracts of the research papers indexed on Web of Science



Source: Author **based on data from https://www.webofscience.com/ (accessed April 2024).**

It should be noted that the words with the highest frequency of appearance are “AI”, “artificial intelligence”, “ethic”, “ethical”, “human”, and “data”, which are closely related to the very criteria according to which data filtering from the Clarivate database is performed. There are also words that reflect the increased level of interest shown for the potential problems related to the use of AI: “risk”, “problem”, “concern”, and “critical”. There are also terms specific to the medical field, which, as was previously mentioned in this work, are linked to many debates on AI ethics, the stake in this case being very important, namely the patient's well-being, “patient”, “healthcare”, “clinical”.

Figure 3. Graphic representation of the evolution of the abstract polarity expressed regarding articles related to AI ethics (published between 1991 and April 2024 and indexed by Web of Science)



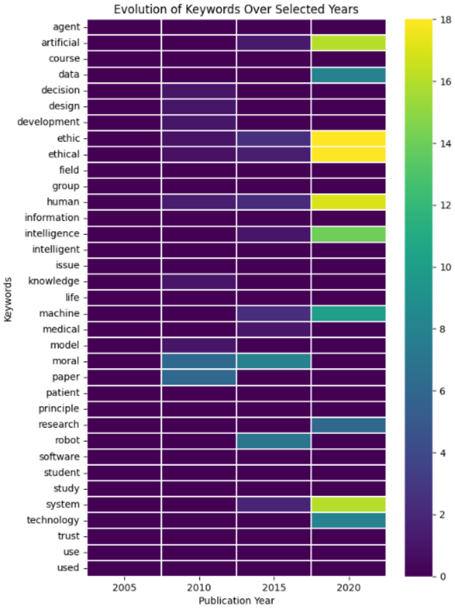
Source: Author.

Sentiment analysis was performed using the Python library called TextBlob, based on the Natural Language Toolkit (NLTK). Subsequently, the Naive Bayes model, implicitly used by TextBlob, was applied to evaluate the sentiments outlined in the abstracts of the selected articles. Subsequently, the polarity is calculated, which can have values ranging from [-1, +1]. The resulting sentiment mean is slightly positive, equal to 0.066, with a standard deviation of 0.087, indicating moderate variation. The minimum value encountered is -0.75, and the maximum value is 0.55. The slightly positive mean value reflects a rather optimistic tone, which can be translated into a vision marked by confidence in the potential of existing AI solutions, technological advancements in the past decade, as well as the potential to increase efficiency levels and reduce the risk of errors. After the analysis, the results showed that 81.5% of the abstracts are positive, 18.1% negative, and 0.4% neutral.

To highlight the evolution of sentiments over the years expressed in scientific works related to AI ethics and trends in the field, a sentiment analysis was carried out depending on the year in which the articles were published. The results can be seen in the figure above (Figure 3). Observing the evolution over the years of the polarity mean calculated through sentiment analysis, we notice that between 2005 and 2010, its value experienced a rather abrupt decrease and came closest to neutrality (value 0), but still remained positive. Behind the sudden drop in the score related to 2007 are multiple significant events. Technological advancement could generate concerns about the ethical aspects of the use of technology. The beginning of the financial crisis could also have led them to adopt a more reserved position and closer to neutrality. It is also notable the progressive decrease in the polarity score that manifests itself starting with the year 2016 and until 2020 The implementation of GDPR in EU member states also brought challenges to companies that were developing AI solutions through stricter requirements regarding confidentiality protection. The years 2020 and 2021, years in which the effects of the COVID-19 pandemic were acutely felt, also had low scores, a sign that this can also be an event with a strong influence on the feelings of the scientific community.

The evolution over the years of the frequency of appearance of keywords in the abstracts of publications included in the study was also analysed.

Figure 4. HeatMap of the evolution of the score of the main keywords



Source: Author.

For this purpose, natural language processing (NLP) techniques such as tokenization and lemmatisation were used. After identification, extraction of keywords, and calculation of appearance frequencies for each word, the Seaborn Python library was used to generate a heatmap graph (Figure 4). It is observed that between the years 2010-2015 there was a greater concern for morality, for what should be considered right or wrong. However, after 2020, concurrently with technological advancements in the field of AI, the incidence of the words "ethical" and "ethic" has increased significantly. Thus, the principles that should govern the field of artificial intelligence and dictate its development directions have acquired multiple nuances, not just black/white, right/wrong, giving rise to many questions that have become intensely debated in specialised literature.

Later, the latent Dirichlet allocation model was used to identify the main topics debated by the academic community regarding the ethics of artificial intelligence. LDA can provide support for a deeper understanding of textual content and for identifying meaningful relationships between words.

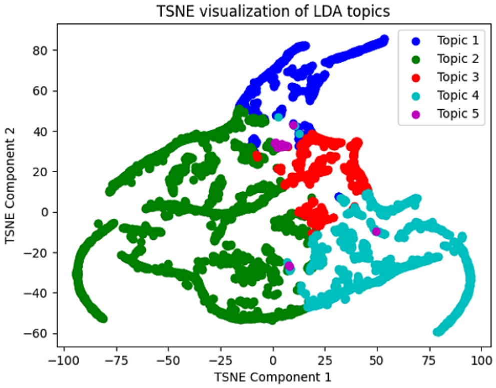
Table 1. The results of applying the LDA model:
The main groups of words and the score of each word

No. of topic	Score*The First Keyword	Score*The Second Keyword	Score*The Third Keyword	Score*The Fourth Keyword
1	0.010*"patient"	0.010*"data"	0.010*"study"	0.009*"method"
2	0.015*"ethical"	0.013*"technology"	0.011*"intelligence"	0.010*"artificial"
3	0.009*"system"	0.008*"bias"	0.007*"human"	0.006*"algorithm"
4	0.020*"human"	0.013*"moral"	0.013*"ethical"	0.012*"system"
5	0.004*"model"	0.004*"urban"	0.004*"case"	0.004*"language"

Source: Author.

In the first topic that results, the presence of the word “patient” is notable, confirming the increased concern about the use of AI solutions in the medical field. The well-being of the patient should be the focus of the attention of specialists. The word “study” reinforces the idea that the analysed documents are an integral part of the specialised literature. The second emerging topic seems to focus on examining the impact of AI solutions' evolution on ethical grounds. The third topic seems to involve arguments and opinions on the shortcomings of machine learning algorithms that underlie artificial intelligence, issues related to bias, and their influence on people. The fourth topic is more diffuse and allows multiple interpretations. It could refer to urban planning, smart cities with technology as a main component, or it could refer to studies based on data from the urban environment.

Figure 5. LDA results displayed by t-SNE



Source: Author.

The graphical representation highlighting the relationships between these 5 main topics was constructed by applying t-distributed Stochastic Neighbour Embedding (t-SNE). Through this method, the dimensionality of the topic distribution obtained by LDA was reduced. In Figure 5 it can be seen that the clusters are generally well defined, but there are also intersection points between them, indicating that there are documents that combine two or more of these topics.

6. Conclusions

After analysing the data set exported in April 2023, which contains data related to 4762 articles in the Web of Science database that refer to the subject of AI ethics, it can be concluded that the sentiments expressed in academic works have a slightly positive sentiment tendency, fuelled by advances in the field. Following the application of the LDA model, it is observed that the subject of AI ethics is a vast one, and the clusters of the main topics addressed are well defined, with their intersections being rare. In the specialised literature, it is noted that there are fields, such as the medical field, where concern about the potential risks and errors of AI is major, and the importance of drawing ethical rules regarding the use of AI solutions is emphasised. Among the limitations of the current study is the fact that the research is based on data extracted from the Clarivate Web of Science platform. Therefore, it is possible that many publications relevant to this topic are not included in this study. At the same time, this study focusses mainly on identifying the main topics and analysing feelings. The method used in the current study for sentiment analysis may encounter difficulties in modelling the nuances of natural language and complex grammatical structures, so in future studies, other methods can be applied to compare the results. In addition, in future work, articles present in other important databases and platforms can be added, and interpretations related to the main ethical problems of AI algorithms described in the specialised literature and the impact they can have on people can be added.

Acknowledgement

This work was supported by a grant from the Ministry of Research, Innovation, and Digitisation, CNCS/CCCDI - UEFISCDI, project number COFUND-DUT-OPEN4CEC-1, within PNCDI IV.

Bibliography

[1] Blanchard, A., Thomas, C., Taddeo, M. (2024). Ethical governance of artificial intelligence for defence: normative tradeoffs for principle to practice guidance. AI and Society. https://doi.org/10.1007/s00146-024-01866-7.

[2] Charan, K.L.S.V.S.N. (2023). Revolutionizing Sentiment Analysis through AI. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 07(09). https://doi.org/10.55041/ijsrem25656.

[3] Chowdhary, K. R. (2020). Fundamentals of artificial intelligence. In Fundamentals of Artificial Intelligence. Springer India. https://doi.org/10.1007/978-81-322-3972-7.

[4] Drabiak, K., Kyzer, S., Nemov, V., El Naqa, I. (2023). AI and machine learning ethics, law, diversity, and global impact. In British Journal of Radiology (Vol. 96, Issue 1150). British Institute of Radiology. https://doi.org/10.1259/bjr.20220934.

[5] Dwivedi, D. N., Mahanty, G., Vemareddy, A. (2022). How Responsible Is AI? International Journal of Information Retrieval Research, 12(1), 1-14. https://doi.org/10.4018/ijirr.298646.

[6] Etzioni, A., Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. Journal of Ethics, 21(4), 403-418. https://doi.org/10.1007/s10892-017-9252-2.

[7] Gao, D. K., Haverly, A., Mittal, S., Wu, J., Chen, J. (2024). A Bibliometric Analysis, Critical Issues, and Key Gaps. International Journal of Business Analytics, 11(1). https://doi.org/10.4018/IJBAN.338367.

[8] Giovanola, B., Tiribelli, S. (2023). Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. AI and Society, 38(2), 549-563. https://doi.org/10.1007/s00146-022-01455-6.

[9] González, A. L., Moreno-Espino, M., Román, A. C. M., Fernández, Y. H., Pérez, N. C. (2024). Ethics in Artificial Intelligence: an Approach to Cybersecurity. Inteligencia Artificial, 27(73), 38-54. https://doi.org/10.4114/intartif.vol27iss73pp38-54.

[10] Hadlington, L., Karanika-Murray, M., Slater, J., Binder, J., Gardner, S., Knight, S. (2024). Public perceptions of the use of artificial intelligence in Defence: a qualitative exploration. AI and Society. https://doi.org/10.1007/s00146-024-01871-w.

[11] Hamilton, R. H., Davison, H. K. (2022). Legal and Ethical Challenges for HR in Machine Learning. Employee Responsibilities and Rights Journal, 34(1), 19-39. https://doi.org/10.1007/s10672-021-09377-z.

[12] Heyen, N. B., Salloch, S. (2021). The ethics of machine learning-based clinical decision support: an analysis through the lens of professionalisation theory. BMC Medical Ethics, 22(1). https://doi.org/10.1186/s12910-021-00679-3.

[13] Karoo, K., Chitte, V. (2023). Ethical Considerations in Sentiment Analysis: Navigating the Complex Landscape. International Research Journal of Modernization in Engineering Technology and Science. https://doi.org/10.56726/irjmets46811.

[14] Leal Filho, W., Ribeiro, P. C. C., Mazutti, J., Lange Salvia, A., Bonato Marcolin, C., Lima Silva Borsatto, J. M., Sharifi, A., Sierra, J., Luetz, J., Pretorius, R., Viera Trevisan, L. (2024). Using artificial intelligence to implement the UN sustainable development goals at higher education institutions. International Journal of Sustainable Development & World Ecology, 1-20. https://doi.org/10.1080/13504509.2024.2327584.

[15] Pool, J., Indulska, M., Sadiq, S. (2024). Large language models and generative AI in telehealth: a responsible use lens. Journal of the American Medical Informatics Association. https://doi.org/10.1093/jamia/ocae035.

[16] Tossell, C. C., Tenhundfeld, N. L., Momen, A., Cooley, K., De Visser, E. J. (2024). Student Perceptions of ChatGPT Use in a College Essay Assignment: Implications for Learning, Grading, and Trust in Artificial Intelligence. IEEE Transactions on Learning Technologies, 17, 1069-1081. https://doi.org/10.1109/TLT.2024.3355015.

[17] Wei, M., Feng, Y., Chen, C., Luo, P., Zuo, C., Meng, L. (2024). Unveiling public perception of AI ethics: an exploration on Wikipedia data. EPJ Data Science, 13(1). https://doi.org/10.1140/epjds/s13688-024-00462-5.

[18] Yoga Saputra, P., Datumaya Wahyudi Sumari, A., Watequlis Syaifudin, Y., Satria Maulana Navalino, V. (2023). Understanding People Opinion on Artificial Intelligence Ethics through Machine Learning-based Sentiment Analysis. In International Journal of Frontier Technology and Engineering (IJFTE) (Vol. 01, Issue 02). http://jurnal.polinema.ac.id/index.php/IJFTE/index.