Migration Letters

Volume: 20, No: S8(2023), pp. 1145-1155

ISSN: 1741-8984 (Print) ISSN: 1741-8992 (Online)
www.migrationletters.com

Natural Language Processing Algorithm for the Classification of Criminal Messages

Kevin Christian Argomedo Pflücker[1], Yenny Milagritos Sifuen- tes Diaz[2], Luis Santiago García Merino[3], Jorge Luis Gutiérrez Gutiérrez[4]

Abstract

A documentary review was carried out on the production and publication of research papers related to the study of the Natural Language Processing Algorithm variable. The purpose of the bibliometric analysis proposed in this document was to know the main characteristics of the volume of publications registered in the Scopus database during the period 2017–2022 by Latin American institutions, achieving the identification of 539 publications. The information provided by this platform was organized through graphs and figures, categorizing the information by Year of Publication, Country of Origin, Area of Knowledge and Type of Publication. Once these characteristics have been described, the position of different authors on the proposed topic is referenced through a qualitative analysis. Among the main findings made through this research, it is found that Brazil with 223 publications with the highest scientific production registered in the name of authors affiliated with institutions in that country. The Area of Knowledge that made the greatest contribution to the construction of bibliographic material related to the study of the Natural Language Processing Algorithm was Computer Science with 450 published documents, and the most used Publication Type during the period indicated above were Conference Articles with 57% of the total scientific production.

Keywords: *Natural Language Processing Algorithm, Criminal Messages.*

1. Introduction

In the era of globalization where the world is increasingly turning to digitalization, the sheer scale of volumes of texts generated every day presents great opportunities and challenges for police department agencies and respective security organizations for different countries. The furor of social networks, online forums and messaging applications has been the bridge in breaking the boundaries in illicit activity, where people with nefarious intentions use these applications to be able to commit crimes through the use of written messages. Detecting and encrypting these messages for criminal purposes has become a priority for law enforcement agencies and police departments in order to minimize cybercrime, terrorism, and extortion.

[1] Ingeniero Informático. Universidad Nacional de Trujillo. Trujillo, Perú, pflucker01@gmail.com, http://orcid.org/0000-0002-0646-4338

[2] Doctora en Educación. Universidad Nacional de Trujillo. Trujillo, Perú, ysifuentes@unitru.edu.pe, http://orcid.org/0000-0001-9464-8294

[3] Doctor en Ingeniería. Universidad Nacional de Trujillo. Trujillo, Perú, jgutierrez@unitru.edu.pe, http://orcid.org/0000-0002-4989-1196

[4] Doctor en Ciencias de la Computación y Sistemas Universidad Católica Los Ángeles de Chimbote Perú, lgarciam@uladech.edu.pe, https://www.orcid.org/0000-0001-9392-2474

The natural language processing formed by artificial intelligence algorithms has played a formidable role in the fight against these types of threats executed for criminal purposes. Natural language processing is a subfield of new technologies that has as its epicenter the interactions that humans have with computers with natural cognitive language. This subfield that offers these cutting-edge technologies allows computers to execute, understand, and interpret human language, providing an invaluable resource for the classification and analysis of textual data. As the prevalence of digital communication grows, so does the need for innovative NLP algorithms that can assist law enforcement and law enforcement agencies in identifying and categorizing criminal messages.

The functionality of this text is to present applications of advanced algorithms offered by natural language processing provided by artificial intelligences, which has been developed in order to classify criminal messages in a faster and more effective way. These security and surveillance tools run by artificial intelligence algorithms have the potential to overpower cybersecurity on a large scale. By harnessing the power of machine learning techniques and cutting-edge linguistic analysis, it aims to equip authorities with the ability to quickly and accurately identify and respond to criminal threats in the digital sphere.

In this comprehensive overview, we'll delve into the various components of these criminal cybersecurity applications, their underlying architecture, and the methods they employ to classify criminal messages. We will also explore its real-world applications, its role in risk mitigation, and the ethical considerations associated with its use.

In this context, the development of these applications underscores the pressing need to bridge the gap between rapidly evolving communication technologies and law enforcement capabilities. As digital communication continues to shape the landscape of criminal activity, it is essential that the authorities in charge of ensuring cybersecurity have the necessary knowledge to be able to quickly combat the threats present in this globalized society. This introduction reflects how a guide is essential to understand and take advantage of the capabilities of these applications offered by artificial intelligence taken hand in hand with natural language processing, which offers a vision of its potential to make our digital world a safer place. For this reason, this article seeks to describe the main characteristics of the compendium of publications indexed in the Scopus database related to the Natural Language Processing Algorithm variables. Such as the description of the position of certain authors affiliated with institutions, during the period between 2017 and 2022.

2. General Objective

To analyze, from a bibliometric and bibliographic perspective, the preparation and publication of research papers in high-impact journals indexed in the Scopus database on the Natural Language Processing Algorithm variable during the period 2017-2022 by Latin American institutions.

3. Methodology

This article is carried out through a research with a mixed orientation that combines the quantitative and qualitative method.

On the one hand, a quantitative analysis of the information selected in Scopus is carried out under a bibliometric approach of the scientific production corresponding to the Natural Language Processing Algorithm study. On the other hand, examples of some research works published in the area of study mentioned above are analyzed from a qualitative perspective, based on a bibliographic approach that allows describing the position of different authors on the proposed topic. It is important to note that the entire

search was carried out through Scopus, managing to establish the parameters referenced in Figure 1.

3.1. Methodological design

Figure 1. Methodological design
Source: Authors' own creation

3.1.1 Phase 1: Data collection

Data collection was carried out from the Search tool on the Scopus website, where 539 publications were obtained from the following filters:

TITLE-ABS-KEY (natural AND language AND processing AND algorithm) AND PUBYEAR > 2016 AND PUBYEAR < 2023 AND (LIMIT-TO (AFFILCOUNTRY , "Brazil") OR LIMIT-TO (AFFILCOUNTRY , "Mexico") OR LIMIT-TO (AFFILCOUNTRY , "Colombia") OR LIMIT-TO (AFFILCOUNTRY , "Ecuador") OR LIMIT-TO (AFFILCOUNTRY , "Chile") OR LIMIT-TO (AFFILCOUNTRY , "Argentina") OR LIMIT-TO (AFFILCOUNTRY , "Peru") OR LIMIT-TO (AFFILCOUNTRY , "Cuba") OR LIMIT-TO (AFFILCOUNTRY , "Uruguay") OR LIMIT-TO (AFFILCOUNTRY , "Costa Rica") OR LIMIT-TO (AFFILCOUNTRY , "Venezuela") OR LIMIT-TO (AFFILCOUNTRY , "Honduras") OR LIMIT-TO (AFFILCOUNTRY , "Puerto Rico") OR LIMIT-TO (AFFILCOUNTRY , "Panama") OR LIMIT-TO (AFFILCOUNTRY , "Guatemala") OR LIMIT-TO (AFFILCOUNTRY , "Bolivia")

- Published documents whose study variables are related to the study of the Natural Language Processing Algorithm variable.
- Limited to the period 2017-2022.
- Limited to Latin American countries.
- Without distinction of area of knowledge.
- No distinction of type of publication.

3.1.2 Phase 2: Construction of analytical material

The information collected in Scopus during the previous phase is organized and then classified by graphs, figures and tables as follows:

- Co-occurrence of words.
- Country of origin of the publication.
- Area of knowledge.
- Type of publication.

3.1.3 Phase 3: Drafting of conclusions and outcome document

In this phase, the results of the previous results are analysed, resulting in the determination of conclusions and, consequently, the obtaining of the final document.

4. Results

4.1 Co-occurrence of words

Figure 2 shows the co-occurrence of keywords found in the publications identified in the Scopus database.

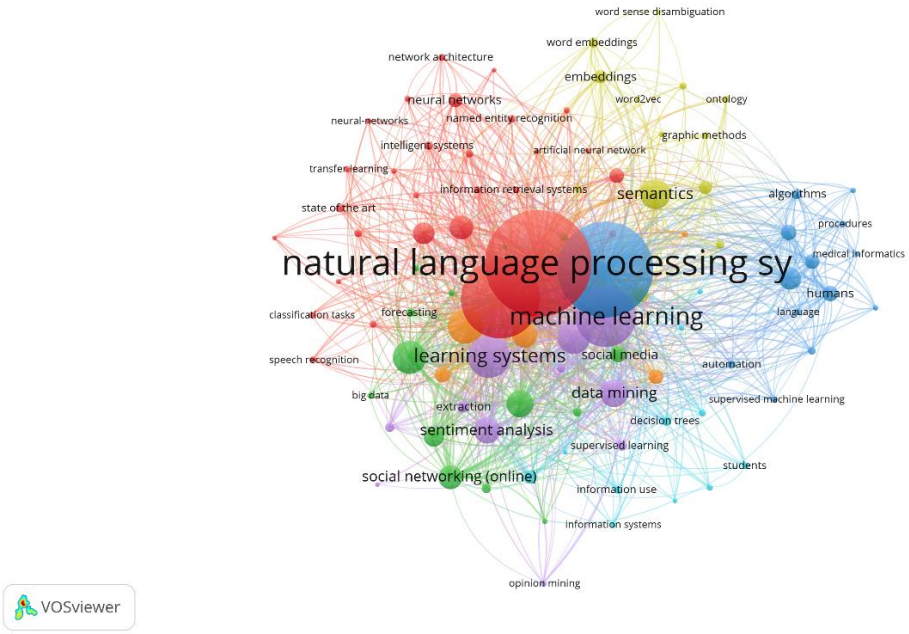

Figure 2. Co-occurrence of words

Source: Authors' own elaboration (2023); based on data exported from Scopus.

Natural Language Processing was the most frequently used keyword within the studies identified through the execution of Phase 1 of the Methodological Design proposed for the development of this article. Learning Systems is among the most frequently used variables, associated with variables such as Sentiment Analysis, Social Networks, Supervised Learning, Semantics, Information Systems, Extraction, Embeddings, Named Entity Recognition. From the above, it is striking, we must always consider the ethical implications of the use of NLP for the classification of criminal messages. It is critical to ensure privacy, prevent bias, and address concerns related to surveillance. Careful consideration of text preprocessing techniques, feature selection, and dimensionality reduction plays a crucial role in the algorithm's performance. Balancing model accuracy and retrieval is essential to minimize false positives and false negatives. Transitioning from a research project to practical implementation in real-world scenarios requires collaboration with law enforcement agencies, legal experts, and policymakers. The algorithm must be integrated into existing systems and comply with relevant regulations and laws.

4.2 Distribution of scientific production by year of publication

Figure 3 shows how scientific production is distributed according to the year of publication.

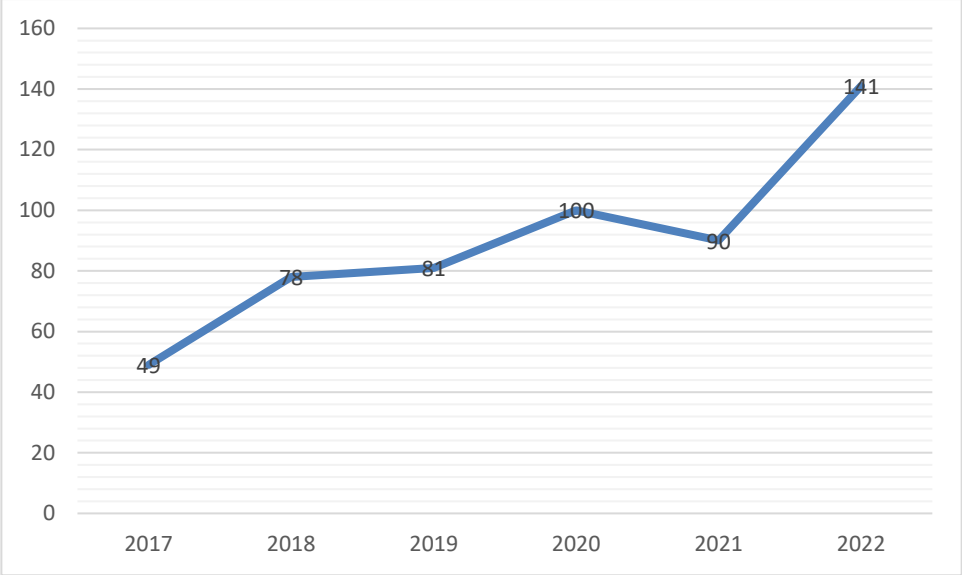

Figure 3. Distribution of scientific production by year of publication.

Source: Authors' own elaboration (2023); based on data exported from Scopus

Among the main characteristics evidenced through the distribution of scientific production by year of publication, the number of publications registered in Scopus was in 2022, reaching a total of 141 documents published in journals indexed on this platform. This article (an extension of our MSR 2020 article) introduces a library called AIMMX for extracting AI model metadata from software repositories to enhanced metadata that conforms to a flexible metadata schema. We evaluated AIMMX against 7998 open-source models from three sources: model zoos, arXiv AI articles, and next-generation AI articles. We also explore how AIMMX can enable studies and tools to advance engineering support for AI development. As preliminary examples, we present an exploratory analysis of the reproducibility of data and methods about the models in the evaluation dataset and a catalog tool for discovering and managing models. We also demonstrate the flexibility of extracted metadata by using the evaluation dataset in an existing natural language processing (NLP) analytics platform to identify trends in the dataset. Overall, we hope that AIMMX will encourage research towards better AI development.(Tsay, 2022)

4.3 Distribution of scientific output by country of origin

Figure 4 shows how scientific production is distributed according to the country of origin of the institutions to which the authors are affiliated.

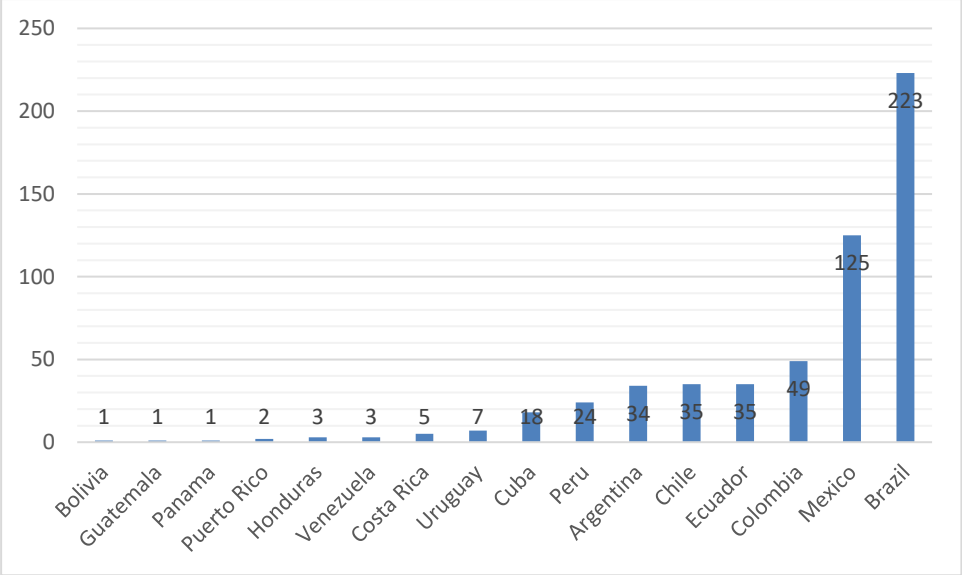

Figure 4. Distribution of scientific production by country of origin.

Source: Authors' own elaboration (2023); based on data provided by Scopus.

Within the distribution of scientific production by country of origin, the registrations from institutions were taken into account, establishing Brazil as the country of this community, with the highest number of publications indexed in Scopus during the period 2017-2022, with a total of 223 publications in total. In second place, Mexico with 125 scientific papers, and Colombia occupying the third place presenting to the scientific community, with a total of 49 documents among which is the article entitled "RastrOS Project: Contributions of Natural Language Processing to the Development of an Eye Tracking Corpus with Predictability Standards for Brazilian Portuguese" This article presents RastrOS, a new eye-tracking corpus of eye movement data from university students during the silent reading of paragraphs of texts in Brazilian Portuguese (BP). The article shows the potential of the corpus for natural language processing (NLP) by using it to evaluate the task of predicting sentence complexity in BP and also focuses on the description of the NLP resources and methods developed to create the corpus. Specifically, we present: (i) the method used to select corpus paragraphs from large corpora, using linguistic metrics and clustering algorithms; (ii) the platform for collecting the Cloze test, which is also responsible for creating the project's datasets, and (iii) the hybrid method of semantic similarity, based on word embedding models and contextualized word representations, used to generate semantic predictability norms. RastrOS can be downloaded from the Open Scientific Framework repository with the computational infrastructure mentioned above. (Leal, 2022)

4.4 Distribution of scientific production by area of knowledge

Figure 5 shows the distribution of the elaboration of scientific publications based on the area of knowledge through which the different research methodologies are implemented.

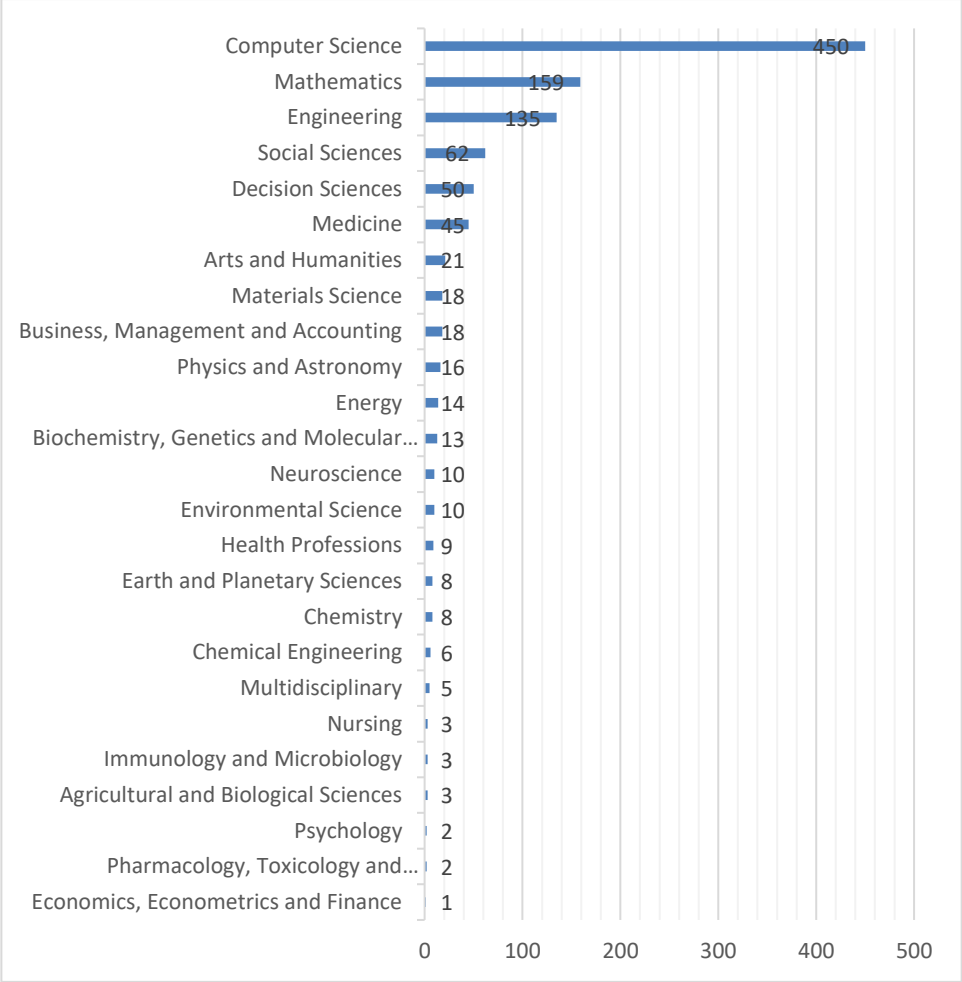

Figure 5. Distribution of scientific production by area of knowledge.

Source: Authors' own elaboration (2023); based on data provided by Scopus

Computer Science was the area of knowledge with the highest number of publications registered in Scopus with a total of 450 documents that have based its methodologies Natural Language Processing Algorithm. In second place, Mathematics with 159 articles and Engineering in third place with 135. The above can be explained thanks to the contribution and study of different branches, the article with the greatest impact was registered by Computer Science entitled "Prediction of interaction design patterns to design explicit interactions in ambient intelligence systems: a case study" the present article aims to present a UIPatternM model to predict interaction design patterns from the processing of text-based requirements through machine learning algorithms. We evaluated the predictions of our proposal. We also present a case study with professional designers who evaluated the UIPatternM recommender's predictions according to a set of design-level requirements that emulate everyday needs. Our participants performed a set of scenario-based tasks and we evaluated participants using effectiveness, efficiency, and satisfaction as performance metrics. The application of the UIPatternM model helped support the conception and refinement of UI design for explicit interaction in AmI systems. (Silva-Rodríguez, 2022)

4.5 Type of publication

In the following graph, you will see the distribution of the bibliographic finding according to the type of publication made by each of the authors found in Scopus.

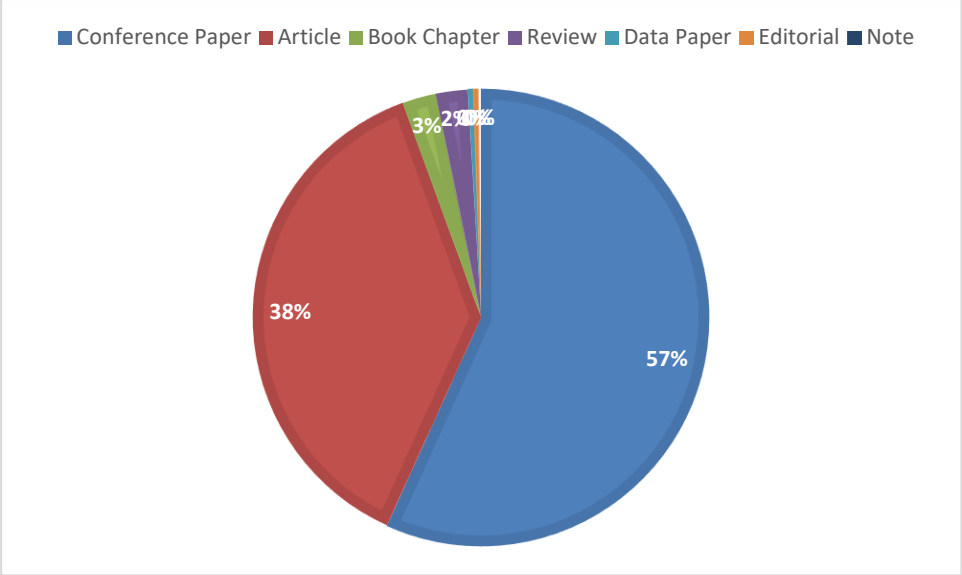

Figure 5. Type of publication.

Source: Authors' own elaboration (2023); based on data provided by Scopus.

The type of publication most frequently used by the researchers referenced in the body of this document was the one entitled Session Paper with 57% of the total production identified for analysis, followed by D Journal Articles with 38%. Chapter of the Book are part of this classification, representing 3% of the research papers published during the period 2017-2022, in journals indexed in Scopus. In this last category, the one entitled "Sociolinguistic repositories as an asset: challenges and difficulties in Brazil" stands out. This article aims to provide a context for documentation in Brazilian Portuguese language and its data collection to establish linguistic repositories from a sociolinguistic perspective. Design/methodology/approach: The main sociolinguistic projects that have generated data collections on the Portuguese language of Brazil are presented. Findings: The comparison with another situation of repositories (seed vaults) and with the accounting concept of assets is evoked to map the challenges to be overcome when proposing a standardized and professional linguistic repository to house the linguistic data collections arising from reported projects and others, in accordance with the principles of the open science movement. Originality/value: Thinking about the sustainability of projects to build linguistic documentation repositories, alliances with the area of information technology, or even with private companies, could minimize the problems of obsolescence and safeguarding of data, by promoting the circulation and automation of analysis through natural language. Processing algorithms. These planning actions can help promote the longevity of the linguistic documentation repositories of Brazilian sociolinguistic research.(Meister Ko. Freitag, 2022)

5. Conclusions

Through the bibliometric analysis carried out in this research work, it was possible to establish that Brazil was the country with the highest number of published records for the Natural Language Processing Algorithm variables. With a total of 223 publications in the Scopus database. In the same way, it was possible to establish that the application of theories framed in the area of Computer Science, were used more frequently in the development of a natural language processing algorithm which has the functionality of being able to encode and classify messages for criminal purposes, which represents an important advance for the field of law enforcement and crime prevention. This technology has the potential to revolutionize the way we identify, track, and respond to criminal activity, making our communities safer. It is important to highlight several

factors that natural language processing presents, which we find the pre-processing of texts executed by artificial intelligence algorithms for the purpose of more automatic learning, this with the creation of a much more solid classification system of messaging applications. The acceptance of the algorithm when it comes to distinguishing the criminal messages executed by cybersecurity has the potential for the police department and the other watchdog to be able to comply with the law and the security professionals to be able to radically combat activities for the purposes of extortion and illegality.

Considering the quality and quantity of data, we have emphasized the importance of high-quality training data and the need for a substantial volume of criminal and non-criminal messages to train an effective model. An extensive, well-labeled dataset is essential for the accuracy and generalizability of the algorithm. Rigorous evaluation methods, such as cross-validation, metrics such as accuracy, precision, retrieval, and F1 score, are critical to evaluating the algorithm's performance and ensuring that it meets the desired goals. Finally, the development and implementation of a natural language processing algorithm for the classification of criminal messages has the potential to improve the efficiency and effectiveness of security and law enforcement efforts. It represents a valuable tool in the fight against crime, while underlining the importance of responsible and ethical use of NLP technology. Further research and collaboration with experts in the field will be crucial to refine and optimize the algorithm for real-world application.

References

Leal, S. E.-G. (2022). RastrOS Project: Contributions of Natural Language Processing to the Development of an Eye-Tracking Corpus with Predictability Standards for Brazilian Portuguese. BRAZIL.

Meister Ko. Freitag, R. (2022). Sociolinguistic repositories as an asset: challenges and difficulties in Brazil. BRAZIL.

Silva-Rodríguez, V. N.-M.-P.-G.-R. (2022). Predicting interaction design patterns for designing explicit interactions in ambient intelligence systems: a case study. MEXICO.

Tsay, J. B. (2022). Extract metadata from enhanced AI models from software repositories. BRAZIL.

Ahn, C. (2023). Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation, 185 doi:10.1016/j.resuscitation.2023.109729

Ajevski, M., Barker, K., Gilbert, A., Hardie, L., & Ryan, F. (2023). ChatGPT and the future of legal education and practice. Law Teacher, doi:10.1080/03069400.2023.2207426

Al Ghatrifi, M. O. M., Al Amairi, J. S. S., & Thottoli, M. M. (2023). Surfing the technology wave: An international perspective on enhancing teaching and learning in accounting. Computers and Education: Artificial Intelligence, 4 doi:10.1016/j.caeai.2023.100144

Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023). ChatGPT: Applications, opportunities, and threats. Paper presented at the 2023 Systems and Information Engineering Design Symposium, SIEDS 2023, 274-279. doi:10.1109/SIEDS58326.2023.10137850 Retrieved from www.scopus.com

Bauer, E., Greisel, M., Kuznetsov, I., Berndt, M., Kollar, I., Dresel, M., . . . Fischer, F. (2023). Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. British Journal of Educational Technology, doi:10.1111/bjet.13336

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. British Journal of Educational Technology, doi:10.1111/bjet.13337

Bender, S. M. (2023). Coexistence and creativity: Screen media education in the age of artificial intelligence content generators. Media Practice and Education, doi:10.1080/25741136.2023.2204203

Berger, U., & Schneider, N. (2023). How will ChatGPT change research, education and healthcare? [Wie wird ChatGPT Forschung, Lehre und Gesundheitsversorgung verändern?] PPmP Psychotherapie Psychosomatik Medizinische Psychologie, 73(3), 159-161. doi:10.1055/A-2017-8471

Busch, F., Adams, L. C., & Bressem, K. K. (2023). Biomedical ethical aspects towards the implementation of artificial intelligence in medical education. Medical Science Educator, doi:10.1007/s40670-023-01815-x

Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. Journal of Medical Systems, 47(1) doi:10.1007/s10916-023-01925-4

Chaudhry, I. S., Sarwary, S. A. M., The Refae, G. A., & Chabchoub, H. (2023). Time to revisit existing Student's performance evaluation approach in higher education sector in a new era of ChatGPT — A case study. Cogent Education, 10(1) doi:10.1080/2331186X.2023.2210461

Choi, E. P. H., Lee, J. J., Ho, M. -., Kwok, J. Y. Y., & Lok, K. Y. W. (2023). Chatting or cheating? the impacts of ChatGPT and other artificial intelligence language models on nurse education. Nurse Education Today, 125 doi:10.1016/j.nedt.2023.105796

Collins, J. E. (2023). Policy solutions: Policy questions for ChatGPT and artificial intelligence. Phi Delta Kappan, 104(7), 60-61. doi:10.1177/00317217231168266

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. Journal of Science Education and Technology, 32(3), 444-452. doi:10.1007/S10956-023-10039-Y

Corsello, A., & Santangelo, A. (2023). May artificial intelligence influence future pediatric research?—The case of ChatGPT. Children, 10(4) doi:10.3390/children10040757

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, doi:10.1080/14703297.2023.2190148

Crawford, J., Cowling, M., & Allen, K. -. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). Journal of University Teaching and Learning Practice, 20(3) doi:10.53761/1.20.3.02

Crawford, J., Cowling, M., Ashton-Hay, S., Kelder, J. -., Middleton, R., & Wilson, G. S. (2023). Artificial intelligence and authorship editor policy: ChatGPT, bard bing AI, and beyond. Journal of University Teaching and Learning Practice, 20(5) doi:10.53761/1.20.5.01

Currie, G. M. (2023). Academic integrity and artificial intelligence: Is ChatGPT hype, hero or heresy? Seminars in Nuclear Medicine, doi:10.1053/j.semnuclmed.2023.04.008

Curtis, N. (2023). To ChatGPT or not to ChatGPT? the impact of artificial intelligence on academic publishing. Pediatric Infectious Disease Journal, 42(4), 275. doi:10.1097/INF.0000000000003852

Dalalah, D., & Dalalah, O. M. A. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. International Journal of Management Education, 21(2) doi:10.1016/j.ijme.2023.100822

Day, T. (2023). A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT. Professional Geographer, doi:10.1080/00330124.2023.2190373

Dergaa, I., Chamari, K., Zmijewski, P., & Saad, H. B. (2023). From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. Biology of Sport, 40(2), 615-622. doi:10.5114/BIOLSPORT.2023.125623

DuBose, J., & Marshall, D. (2023). AI in academic writing: Tool or invader. Public Services Quarterly, 19(2), 125-130. doi:10.1080/15228959.2023.2185338

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., . . . Wright, R. (2023). "So what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71 doi:10.1016/j.ijinfomgt.2023.102642

Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. Journal of University Teaching and Learning Practice, 20(5) doi:10.53761/1.20.5.02

Eggmann, F., Weiger, R., Zitzmann, N. U., & Blatz, M. B. (2023). Implications of large language models such as ChatGPT for dental medicine. Journal of Esthetic and Restorative Dentistry, doi:10.1111/jerd.13046

Ellaway, R. H., & Tolsgaard, M. (2023). Artificial scholarship: LLMs in health professions education research. Advances in Health Sciences Education, doi:10.1007/s10459-023-10257-4

Emenike, M. E., & Emenike, B. U. (2023). Was this title generated by ChatGPT? considerations for artificial intelligence text-generation software programs for chemists and chemistry educators. Journal of Chemical Education, 100(4), 1413-1418. doi:10.1021/acs.jchemed.3c00063