The Effect of Natural Language Processing on the Analysis of Unstructured Text: A Systematic Review

Walter Luis Roldan-Baluis, Noel Alcas Zapata, Maria Soledad Mañaccasa Vásquez
Postgraduate School, Doctorate in Education
Cesar Vallejo University, Lima, Perú

***Abstract*—The analysis of the unstructured text has become a challenge for the community dedicated to natural language processing (NLP) and Machine Learning (ML). This paper aims to describe the potential of the most used NLP techniques and ML algorithms to address various problems afflicting our society. Several original articles were reviewed and published in SCOPUS during 2021. The applied approach was retrospective, transversal and descriptive. The data collected were entered into the SPSS statistical software v25 and among the findings, it was determined that the most used NLP technique was the Term frequency - Inverse document frequency (TF-IDF), while the most used supervised learning algorithm was the Support Vector Machines (SVM). Likewise, the predominant deep learning algorithm was Long Short-Term Memory (LSTM). This research aims to support experts and those starting in research to identify the most used algorithms of NLP and ML.**

Keywords*—*Artificial intelligence; natural language processing; machine learning; unstructured text analysis

I. INTRODUCTION

The Internet has become an exclusive ally for any institution. According [1], there were more than 5'168,000,000 users worldwide, of which Asia accounts for 53.4%, and ranks first. Latin America and the Caribbean are positioned in ranks fourth with 9.6%. According to [2], in Spain the number of users reached 91% of the population. The author in [3] points out that there are currently more than 1'900,000,000 web pages. It also points out that 167 million videos are generated in a minute on the Tik Tok platform; likewise, Amazon customers invest USD 283,000 in e-commerce. It is concluded that in every second large amounts of data are produced in various formats such as images, audios, videos and texts.

The massive amount of data implies the need to automate human tasks through the fast advance of technological innovation. It can be used for decision making in an efficient and effective way. According to [4], such innovation includes Artificial Intelligence (AI). The author in [5] points out that AI trains computers to learn from experience and to do the work of human beings. This field has had an intensified growth due to the COVID-19 pandemic. In the research [6] states that AI surpasses the cognitive abilities of man. AI is an interdisciplinary field [7], of computing capable of solving problems of medicine, psychology, education, health, information technologies - TIC, among others.

On the Internet platform, there is a lot of traffic, users are producing a high volume of unstructured texts; it is difficult to determine which websites are visited by users. The author in [8] propose a model made based on Natural Language Processing (NLP) techniques and neural networks to identify the websites visited by users by translating this problem into a text classification context. This solution is advantageous for the digital marketing because it allows the loyalty of users.

Social networks are platforms on which there are a high proliferation of comments, with absolute freedom and without restrictions from attacks, insults, discriminatory speeches, hatreds and other offensive terms. In the research work [9] proposes a text classification model to detect cyberbullying consisting of a neural network framework that examines the content of the text in order to analyze the effect of the extracted characteristics. The usefulness of this study lies in identifying solid mechanisms for the detection of cyberbullying.

Regarding the scope of the research, systematic review articles published in the SCOPUS Database, period 2021, were reviewed. For example, [10] submitted a systematic review article to provide evidence on the properties of text data used to train machine learning approaches and how they can be applied in clinical practice. In another review article, [11] highlighted the usefulness of NLP and Machine Learning (ML) to structure the comments of free texts issued by patients of health organizations. This led to identify that there is no systematic research that describes the frequency of the ML algorithms used in the various original articles. This work aims to fill this gap, being this the main motivation to carry out this research.

The objective of this study is to systematically review the bibliography of the application of natural language processing to analyze, interpret and classify the high production of unstructured texts produced in digital format. Also, to describe the frequency of ML algorithms such as supervised, unsupervised and deep learning. In this context, the aim is to answer the questions raised in Table I. Question one aims to identify NLP application fields. These features include text preprocessing techniques such as tokenization, etc. Question two describes the frequency of ML algorithms for data analysis. Finally, question three refers to the frequency of deep learning algorithms; this question has been given preference since algorithms are very specialized.

TABLE I. RESEARCH QUESTIONS

No.	Question
1	What are NLP and ML application fields?
2	What is the frequency of ML algorithms for text analysis?
3	What is the frequency of DL algorithms?

II. Literature Review

Natural Language Processing – NLP is a branch of artificial intelligence and a resource to carry out qualitative tasks of unstructured information, based on mathematical and statistical algorithms on large amounts of data. In this regard, [11] pointed out that the NLP is a computer analytical technique used to extract information from an unstructured text into a structured form, for which syntactic processing of a text is done; it also captures the meaning and identifies links based on semantic relationships. The author in [12] indicates that NLP is a technique of automatic extraction of information from different electronically written resources at the level of documents, words, grammar, meaning, and context. Likewise, [13] stated that the NLP is a key tool for information automation and extraction that can process large amounts of data and its application is useful for issuing reports from the radiology area of a hospital.

Machine Learning (ML) is used for creating models that allow computers to learn without being programmed. In this regard, [11] affirm that ML is a set of statistical algorithms that can train and test a group of data to detect patterns, predict feelings within a text. In the research [14] point out that ML is the process that detects and exploits patterns and trends that are "hidden" in the production of unstructured texts. The author in [15] indicate that ML is a set of machine learning algorithms built into machines to provide knowledge about processes quickly and efficiently. ML is classified into three fields, supervised learning (S), unsupervised learning (US), and reinforcement learning. The present work contemplates the use of algorithms of the first two fields mentioned.

Supervised learning uses algorithms that learn iteratively from data. They find hidden information by which computers learn. The author in [11] point out that algorithms try to predict and classify texts. For example, in the electronic documentation of a health service, the algorithms are able to identify the most common issues expressed by patients. Likewise, [16] points out that supervised ML divides the input data set into training and testing. The training data set has an output variable that must be predicted or classified.

Unsupervised learning uses algorithms to identify patterns and detect anomalies such as fraud, scam of potential users, among others. In this regard, [11] indicates that it is a technique that identifies models or patterns of behavior without the need to know the target attribute or objective that could be present in a text. In the research work [16] points out that the algorithms learn some characteristics from the data. One of the best-known models is the clustering.

The NLP is taking a lot of relevance in the sentiment analysis (SA), positive or negative, in the analysis of unstructured text; a source of application is the comments that are made on social networks. In that aspect, [17], making use of the tasks of NLP and ML methods, propose a model of word processing for SA that uses the comments made on Twitter. The first phase consists of collecting the text, cleaning it, preprocessing, extracting features from a text and then categorizing the data. The proposed corpus is multidisciplinary and can be used in the area of market analysis, customer behavior, survey analysis, and brand monitoring, among others. This contribution is used as a basis for broadening the range of real applications.

The usefulness of NLP and ML has a high level of application in the medicine field. It can be applied to determine the misuse and abuse of prescription drugs in comments made on social networks. In this regard, [18] propose a model to detect self-reports of prescription drug abuse from Twitter. Using these public data, it develops a continuous monitoring system to classify the class of "abuse or misuse".

III. Material and Method

A. Introduction

The PRISMA method is a structured tool with a systemic approach that helps to present the results of a research. According to [19], the Preferred Reporting Items for Systematic Reviews and Meta-Analyses – PRISMA 2020 is conceptualized as a series of recommendations that contribute to selecting, evaluating and synthesizing for better clarity and transparency of research. In fact, [20], [21] point out that the PRISMA declaration is an essential strategy for conducting good research and publishing the results. In the area of objectivity, this research has been divided into four phases, according to the process proposed by:

1) Retrieval of publications.
2) Review of titles and abstracts.
3) Revision of the full text.
4) System information collection.

With regard to the initial recovery phase, it is necessary to use a strategy that would allow efficient document searches. In this respect, [22] point out that the PICO strategy is relevant for raising research questions in order to optimize the placement of articles. The PICO system is an acronym and a component structure. According to [23], this format has four elements: problem, intervention, comparison and outcome. Table II shows the optimal search of documents, this strategy was adapted to the acronym PIO. In addition, the thesaurus Computer Classification System – ACM was used to identify the appropriate synonyms; the link is: https://bit.ly/3dphAJP.

From phase two: review, titles and abstracts; articles were located to be contrasted with the inclusion and exclusion criteria. Titles and abstracts were reviewed, then the method and results, in order to establish the search formula. The database consulted was Scopus, period 2021. In the third phase, the combination of keywords and synonyms was used with emphasis on the variables Natural Language Processing and text analysis. The logical operators AND and OR were used repeatedly until the appropriate formula was obtained. Table III shows the restricted query.

TABLE II. Keywords and Synonyms for the PIO Method Search

P	I	O
Natural Language Processing Natural language process Natural Language Text Computational linguistics Word processing NLP	text analysis text analytics text data	Corpus Classification

TABLE III. SEARCH CRITERIA FOR ORIGINAL SCIENTIFIC ARTICLES

Database	Search date	Search string
Scopus	December 15, 2021	OA(all) AND (TITLE-ABS("Natural Language Processing") OR TITLE-ABS("Natural Language Process") OR TITLE-ABS("Natural Language Text") OR TITLE-ABS("Computational linguistics") OR TITLE-ABS("Word processing") OR TITLE-ABS("NLP")) AND (TITLE-ABS("text analysis") OR TITLE-ABS("text analytics") OR TITLE-ABS("text data") OR TITLE-ABS("text classification") OR TITLE-ABS ("Data extraction")) AND PUBYEAR > 2020 AND DOCTYPE(AR)

B. Selection of Criteria

Inclusion and exclusion criteria for the efficient search of research articles were identified in the PICO strategy. The query was held on December 19, 2021. The search was restricted since 2021 and 144 articles were located in *Scopus* database. To ensure the rigor and credibility of the selected articles, they were evaluated by extrapolating the criteria defined in Table IV.

TABLE IV. INCLUSION AND EXCLUSION CRITERIA USING THE PICOS MODEL

PICOS	Inclusion criteria	Exclusion criteria
Problem	Natural language processing – NLP in the text analysis	Natural language processing in formats other than texts (e.g., video, audio).
Intervention	NLP interventions in the data extraction and summaries of text analysis with free software (R language, Python)	NLP interventions in which actual text, using an NLP process, is not processed. Data extraction with licensed software. *Chatbot.*
Comparison	Comparison with other type of intervention such as the elaboration of the linguistic corpus.	Studies that have no other type of comparison.
Outcomes	Report on the impact of the intervention.	It does not contain a report on the impact of the intervention.
Study Type	Quantitative, qualitative, and mixed method studies of original articles.	Systematic review articles, meta-analysis, literature reviews, conferences, dissertations, protocol works, tutorials. Studies not conducted in English. Duplicate jobs and not available in full text.

IV. RESULTS

A. Search Results

A total of 144 articles were collected during the search process. 11 were deleted after reviewing the title and abstract (n=133) of each document. Then, the method and conclusions were reviewed with emphasis and those that did not meet the inclusion criteria were discarded (n=87). Finally, there were 46 potential articles for systematic review. Fig. 1 shows the flowchart of the search strategy.

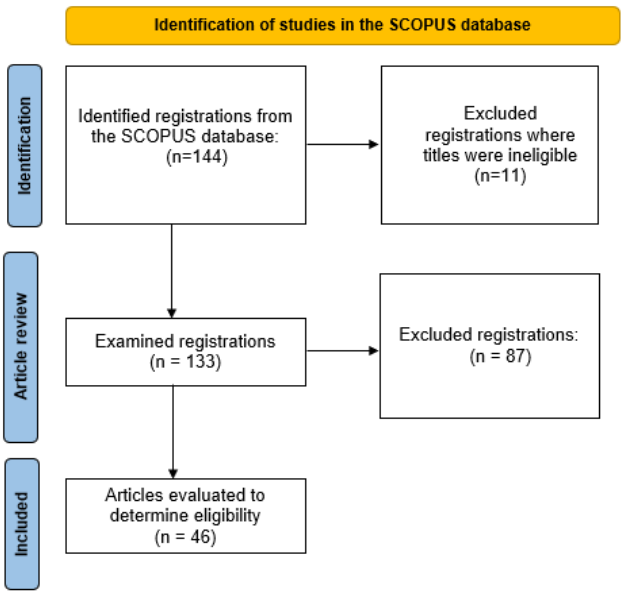

Fig. 1. Flowchart of the Literature Review Process.

B. Description of Included Studies

The application fields or sectors that have benefited from the NLP and ML application correspond to the domains such as aviation, medicine, cyberbullying, education, engineering, technology, among others. In this regard, the medical sector has benefited from 13 studies, 28.76%. The education system from 10 studies, 21.74%. The Technology field has seven articles, 15.22%, among others. Table V shows the details.

TABLE V. FIELDS OF APPLICATION AND FREQUENCY OF ARTICLES

Application field	Frequency	%	Author(s)
Aviation	1	2.17	[24]
Natural disaster relief	1	2.17	[25]
Cyberbullying	1	2.17	[9]
Construction	1	2.17	[26]–[27]
Software Development	3	6.52	[28]–[30]
Education	10	21.74	[31]–[40]
Finance	1	2.17	[41]
Engineering	1	2.17	[42]
Marketing	2	4.35	[8]–[17]
Medicine	13	28.26	[43]–[55]
Business organization	1	2.17	[56]
Politics	1	2.17	[57]
Psychology	1	2.17	[58]
Information security	1	2.17	[59]
Technology	7	15.22	[60]–[66]
Urban transport	1	2.17	[67]

C. Frequency of NLP Algorithms

A number of NLP techniques were applied prior to the use of ML algorithms, and these NLP techniques are described in Appendix A. NLP techniques are associated with various algorithms, which are defined in Appendix B. After registering the data in the SPSS v25 statistical software, it has been discovered that the Term frequency - inverse document frequency (TF-IDF) algorithm was present in 22 articles, 47.82%. The Word2Vec algorithm was used 15 times, 32.60%. *Glove* algorithm at 14, 30.43%, while Bag of words (*BoW)* was used in 11 studies, 23.91%, and N-Gram was used in six articles, 13.04%.

On the other hand, seven studies, 15.21%, used three algorithms at the same time. Likewise, 12 articles, 26.09%, used two algorithms at the same time, while 26 articles used a single algorithm, 58.69%, in their research. The details of the NLP algorithms used are detailed in Appendix C.

D. Frequency of ML Algorithms

ML algorithms analyzed in this study are defined in Appendix D and grouped under supervised learning, unsupervised learning, and Deep Learning.

1) Frequency of supervised learning algorithms (S): The Support Vector Machine (SVM) algorithm was used in 17 studies. The Naive Bayes (NB) algorithm was applied in 15 studies. The Radom Forest (RF) algorithm has 10 studies. R has 9 studies. K-NN has 8 studies. RF has 5 studies. The Passive aggressive (PA) algorithm was used in two studies. The AdaBoost (ADA) algorithm has 1 study like Singular Value Decomposition (SVD) algorithm, Fig. 2 shows the details.

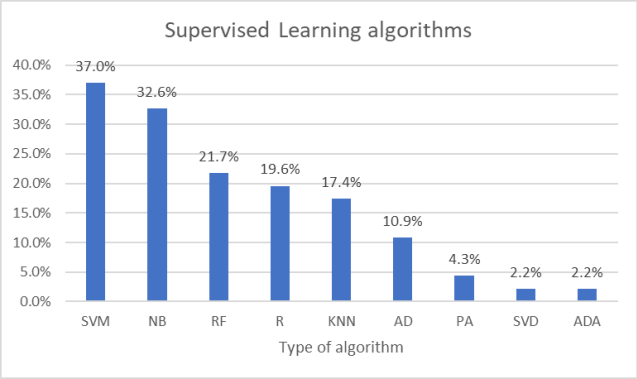

Fig. 2. Studies using S Algorithms.

2) Frequency of unsupervised learning algorithms (US): The Latent Dirichlet Allocation (LDA) algorithm was used in three studies, 6.5%, while K-Means algorithm was used only in one study, 2.2%.

3) Frequency of deep learning algorithms (DL): The Long Short Term Memory (LSTM) deep learning algorithm has 24 studies. Then, the Convolutional neural networks (CNN) algorithm has 12 studies. The Recurrent Neural Networks (RNN) algorithm has 9 studies. The Multilayer Perceptron (MLP) algorithm has 5 studies. The least used algorithms were Gating Circulation Unit (GRU) algorithm with four studies and Artificial neural network (ANN) with two studies. Fig. 3 shows the details.

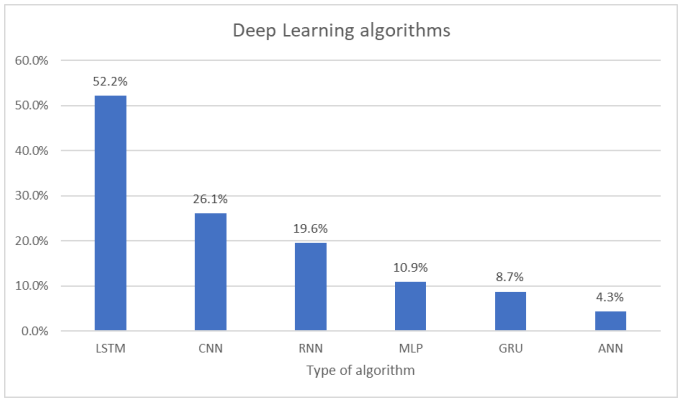

Fig. 3. Studies using DL Algorithms.

4) Studies with hybrid algorithms, Supervised (S), Unsupervised (US) and Deep Learning (DL): Out of 46 articles, 10 (21.74%) use only S algorithms. In this regard, [24] and [52] use SVM. The author in [35] use NB algorithm. On the other hand, the study of [51] uses the LDA algorithm.

Three studies use S and US algorithms at the same time: [17] use two S algorithms: SVM, NB and 1 NS: K-Means. [67] use five S algorithms: SVM, NB, Regression (R), RF, KNN, and one US algorithm: LDA.

The author in [25] uses three algorithms: A supervised learning algorithm, the LDA, and two deep learning algorithms, GRU and CNN. Twenty-one studies, 45.65%, used only DL algorithms, in particular [8], [32], [48] y [68].

Eleven studies, 23.91%, used S and DL algorithms at the same time, in particular [9], [36], [38], [46], [47], [49], [53], [55], [57], [58] y [60].

The details of the ML algorithms used by the 46 studies can be found in Appendix C.

V. Discussion

The popularity and proliferation of platforms working on the Internet such as web portals, social networks, and all digital media have created a massive social interaction between users, even more so because of the global COVID-19 pandemic that has led to the unprecedented increase in online learning and its consequent exponential production of unstructured texts. This phenomenon, according to [32], is allowing the increasing use of the NLP and ML field in text analysis for an efficient solution of real problems.

What are NLP and ML application fields?

It was discovered that sectors such as the health system, education, technology, engineering, software development, aviation, natural disasters relief, cyberbullying, construction, finance, marketing, politics, business organization, information security, psychology, and urban transport, benefit most. This reflects that NLP and ML can be applied to solve problems in any sector.

What is the frequency of ML algorithms for text analysis?

The analysis of the articles indicates that NLP preprocessing techniques such as tokenization, normalization, elimination of irrelevant words are necessary to apply ML algorithms, which allow having a positive impact to achieve the Garg & Sharma study objective [17]. TF-IDF, word2Vec, and Glove are among the most used NLP algorithms. The ML algorithms of supervised learning were: SVM, NB and RF. The least used algorithms were: PA, ADA and SVD. With respect to unsupervised learning algorithms, these were the least used. Only three studies used the LDA algorithm.

What is the frequency of DL algorithms?

With regard to Deep learning, the most used algorithm was LSTM with 22 articles, and the least used was ANN with only 2 articles. This approach becomes a primary tool for the NLP. However, it should be noted that ML algorithms can lead to error bias because it depends on the quality of data with which the research is carried out and especially on access to data since many institutions, unfortunately, restrict them, for example, hospitals [69].

Considering the works obtained, it can be said that the most used NLP technique was the TF-IDF. The most used supervised learning algorithm was the SVM, and with respect to neural networks or deep learning, it was the LSTM. On the other hand, according to [69], one of the main obstacles to applying the NLP and ML algorithms is access to data, representing a challenge for the AI community in reversing this situation.

VI. Conclusion

The most commonly used supervised learning algorithms for text analysis in the field of research are TF-IDF, word2Vec and Glove, while predominant deep learning algorithms are LSTM and ANN. In addition, this article complements the various studies regarding systematic reviews on NLP and ML, by describing the frequencies of influential algorithms and it is expected that this work will lead to further research to increase the cases of application of PLN and ML for the benefit of various fields such as health, education, transport, technology and others. Finally, it should be noted that improving the cognitive aspect of this science requires further research taking into account that the PLN and ML algorithms are universal, characteristic of mathematics and statistics.

References

[1] Internet World Stats, Global Threat Report 2021, 1d. C., 2022, https://www.internetworldstats.com/stats.htm.

[2] S. Galeano, M4rketing Ecommerce, Qué pasa en Internet en un minuto en 2021, 2022, https://marketing4ecommerce.net/que-pasa-en-internet-en-un-minuto-infografia/.

[3] Internet live stat, Total number of Websites, 2021. https://www.internetlivestats.com/total-number-of-websites/.

[4] D. Gruson, «Big Data , inteligencia artificial y medicina de laboratorio : la hora de la integración», Adv Lab Med, 2(1), 5-7, 2021, https://doi.org/10.1515/almed-2021-0014.

[5] H.P. Winston, Artificial Inteligence, Third, Enited States of America, 1992. https://courses.csail.mit.edu/6.034f/ai3/rest.pdf.

[6] D.F. Arbeláez-Campillo, J.J. EspinozaV illasmil, M.J. Rojas-Bahamón, «Inteligencia artificial y condición humana: ¿Entidades contrapuestas o fuerzas complementarias?», Revista de ciencias sociales, 27(2), 502-513, 2021, doi:10.31876/rcs.v27i2.35937.

[7] D. Garabato, Análisis no supervisado de observaciones atípicas en la misión espacial Gaia: optimización mediante procesamiento distribuido e integración en Apsis, Universidade da Coruña, 2020. https://ruc.udc.es/dspace/handle/2183/26479.

[8] D. Perdices, J. Ramos, J.L. García-Dorado, I. González, J.E. López de Vergara, «Natural language processing for web browsing analytics: Challenges, lessons learned, and opportunities», Computer Networks, 198, 2-14, 2021, doi:10.1016/j.comnet.2021.108357.

[9] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, M. Prasad, «Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques», Electronics (Switzerland), 10(22), 1-20, 2021, doi:10.3390/electronics10222810.

[10] I. Spasic, G. Nenadic, «Clinical text data in machine learning: Systematic review», JMIR Medical Informatics, 8(3), 2020, doi:10.2196/17984.

[11] M. Khanbhai, P. Anyadi, J. Symons, K. Flott, A. Darzi, E. Mayer, «Applying natural language processing and machine learning techniques to patient experience feedback: A systematic review», BMJ Health and Care Informatics, 28(1), 1-12, 2021, 10.1136/bmjhci-2020-100262.

[12] S.R. Jonnalagadda, P. Goyal, M.D. Huffman, «Automating data extraction in systematic reviews: A systematic review», Systematic Reviews, 4(1), 2015, doi:10.1186/s13643-015-0066-7.

[13] E. Pons, M. Loes, M. Braun, M. Hunink, J. Kors, «natural Language Processing in Radiology: A Systematic Review1», Radiology, 279(2), 329-343, 2021, doi:10.1148/radiol.16142770.

[14] M. Ceriotti, C. Clementi, O. Anatole Von Lilienfeld, «Introduction: Machine Learning at the Atomic Scale», Chemical Reviews, 121(16), 9719-9721, 2021, doi:10.1021/acs.chemrev.1c00598.

[15] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, A. Aljaaf, Springer Link, A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science, 2019.

[16] M. Batta, «Machine Learning Algorithms - A Review», International Journal of Science and Research (IJ, 9(1), 381-386, 2020, doi:10.21275/ART20203995.

[17] N. Garg, K. Sharma, «Text pre-processing of multilingual for sentiment analysis based on social network data», International Journal of Electrical and Computer Engineering, 12(1), 776-784, 2022, doi:10.11591/ijece.v12i1.pp776-784.

[18] M.A. Al-Garadi, Y.C. Yang, H. Cai, Y. Ruan, K. O'Connor, G.H. Graciela, J. Perrone, A. Sarker, «Text classification models for the automatic detection of nonmedical prescription medication use from social media», BMC Medical Informatics and Decision Making, 21(1), 1-13, 2021, doi:10.1186/s12911-021-01394-0.

[19] D. Moher, «Reporting guidelines: Doing better for readers», BMC Medicine, 16(1), 18-20, 2018, doi:10.1186/s12916-018-1226-0.

[20] R. Sarkis-Onofre, F. Catalá-López, E. Aromataris, C. Lockwood, «How to properly use the PRISMA Statement», Systematic Reviews, 10(1), 13-15, 2021, doi:10.1186/s13643-021-01671-z.

[21] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S.F. Jones, R. Forshee, M. Walderhaug, T. Botsis, «Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review», Journal of Biomedical Informatics, 73, 14-29, 2017, doi:10.1016/j.jbi.2017.07.012.

[22] E. Landa-Ramírez, A. de J. Arredondo-Pantaleón, «Herramienta pico para la formulación y búsqueda de preguntas clínicamente relevantes en la psicooncología basada en la evidencia», Psicooncología, 11(2-3), 250-270, 2014, doi:10.5209/rev_PSIC.2014.v11.n2-3.47387.

[23] A. Pérez Ortiz, M. Ortega Luyando, A. Amaya Hernández, «Programas de prevención de obesidad infantil en México: una revisión sistemática PICO», Psicología y Salud, 31(2), 169-177, 2021, doi:10.25009/pys.v31i2.2686.

[24] T. Madeira, R. Melício, D. Valério, L. Santos, «Machine learning and natural language processing for prediction of human factors in aviation incident reports», Aerospace, 8(2), 1-18, 2021, doi:10.3390/aerospace8020047.

[25] S. Khatoon, M.A. Alshamari, A. Asif, M.M. Hasan, S. Abdou, K.M. Elsayed, M. Rashwan, «Development of social media analytics system for emergency event detection and crisismanagement», Computers, Materials and Continua, 68(3), 3079-3100, 2021, doi:10.32604/cmc.2021.017371.

[26] M.L. Yu, M.H. Tsai, «ACS: Construction data auto-correction system-Taiwan public construction data example», Sustainability (Switzerland), 13(1), 1-21, 2021, doi:10.3390/su13010362.

[27] S. Moon, G. Lee, S. Chi, H. Oh, «Automated Construction Specification Review with Named Entity Recognition Using Natural Language Processing», Journal of Construction Engineering and Management, 147(1), 1-12, 2021, doi:10.1061/(asce)co.1943-7862.0001953.

[28] Ş. Ozan, «Case studies on using natural language processing techniques in customer relationship management software», Journal of Intelligent Information Systems, 56(2), 233-253, 2021, doi:10.1007/s10844-020-00619-4.

[29] M. Alenezi, Z. Mohammed, Y. Javed, «Efficient deep features learning for vulnerability detection using character ngram embedding», Jordanian Journal of Computers and Information Technology, 7(1), 25-38, 2021, doi:jjcit.71-1597824949.

[30] P. Rani, S. Panichella, M. Leuenberger, A. Di-Sorbo, O. Nierstrasz, «How to identify class comment types? A multi-language approach for class comment classification», Journal of Systems and Software, 181, 2-17, 2021, doi:10.1016/j.jss.2021.111047.

[31] L. Burdick, J.K. Kummerfeld, R. Mihalcea, «To batch or not to batch? Comparing batching and curriculum learning strategies across tasks and datasets», Mathematics, 9(18), 2021, doi:10.3390/math9182234.

[32] M. Mujahid, E. Lee, F. Rustam, P.B. Washington, S. Ullah, A.A. Reshi, I. Ashraf, «Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19», Applied Sciences, 11(18), 1-25, 2021, doi:10.3390/app11188438.

[33] R. Adipradana, B.P. Nayoga, R. Suryadi, D. Suhartono, «Hoax analyzer for Indonesian news using RNNs with fasttext and glove embeddings», Bulletin of Electrical Engineering and Informatics, 10(4), 2130-2136, 2021, doi:10.11591/eei.v10i4.2956.

[34] O.C. Stringham, S. Moncayo, K.G.W. Hill, A. Toomes, L. Mitchell, J. V. Ross, P. Cassey, «Text classification to streamline online wildlife trade analyses», PLOS ONE, 16(7), e0254007, 2021, doi:10.1371/journal.pone.0254007.

[35] S.G. Chethan, S. Vinay, «Analytical Framework for Binarized Response for Enhancing Knowledge Delivery System», International Journal of Advanced Computer Science and Applications, 12(10), 348-358, 2021, doi:10.14569/ijacsa.2021.0121157.

[36] R.A. Farouk, M.H. Khafagy, M. Ali, K. Munir, R. M.Badry, «Arabic Semantic Similarity Approach for Farmers‟ Complaints», International Journal of Advanced Computer Science and Applications, 12(10), 348-358, 2021, doi:10.14569/IJACSA.2021.0121038.

[37] C.X. Zhang, R. Liu, X.Y. Gao, B. Yu, «Graph Convolutional Network for Word Sense Disambiguation», Discrete Dynamics in Nature and Society, 2021, 1-12, 2021, doi:10.1155/2021/2822126.

[38] J.L. Huan, A.A. Sekh, C. Quek, D.K. Prasad, «Emotionally charged text classification with deep learning and sentiment semantic», Neural Computing and Applications, 1, 1-11, 2021, doi:10.1007/s00521-021-06542-1.

[39] H.X. Huynh, L.X. Dang, N. Duong-Trung, C.T. Phan, «Vietnamese Short Text Classification via Distributed Computation», International Journal of Advanced Computer Science and Applications, 12(7), 23-31, 2021, doi:10.14569/IJACSA.2021.0120703.

[40] Y. Hu, H. Shen, W. Liu, F. Min, X. Qiao, K. Jin, «A Graph Convolutional Network with Multiple Dependency Representations for Relation Extraction», IEEE Access, 9, 1-14, 2021, doi:10.1109/ACCESS.2021.3086480.

[41] D. Alsaleh, S. Larabi-Marie-Sainte, «Arabic Text Classification Using Convolutional Neural Network and Genetic Algorithms», IEEE Access, 9, 91670-91685, 2021, doi:10.1109/ACCESS.2021.3091376.

[42] S. Sarica, J. Luo, «Stopwords in technical language processing», Plos One, 16(8), 1-13, 2021, doi:10.1371/journal.pone.0254937.

[43] C.J. Harrison, C.J. Sidey-Gibbons, «Machine learning in medicine: a practical introduction to natural language processing», BMC Medical Research Methodology, 21(1), 1-11, 2021, doi:10.1186/s12874-021-01347-1.

[44] P. López-Úbeda, A. Pomares-Quimbaya, M.C. Díaz-Galiano, S. Schulz, «Collecting specialty-related medical terms: Development and evaluation of a resource for Spanish», BMC Medical Informatics and Decision Making, 21(1), 1-17, 2021, doi:10.1186/s12911-021-01495-w.

[45] W. Wang, A. Feng, «Self-Information Loss Compensation Learning for Machine-Generated Text Detection», Mathematical Problems in Engineering, 2021, 1-7, 2021, doi:10.1155/2021/6669468.

[46] A.T. Bako, H.L. Taylor, K. Wiley, J. Zheng, H. Walter-McCabe, S.N. Kasthurirathne, J.R. Vest, «Using natural language processing to classify social work interventions», American Journal of Managed Care, 27(1), 1-18, 2021, doi:10.37765/AJMC.2021.88580.

[47] V. Kumar, D.R. Recupero, D. Riboni, R. Helaoui, «Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification from Clinical Notes», IEEE Access, 9(2021), 7107-7126, 2021, doi:10.1109/ACCESS.2020.3043221.

[48] D. Rakesh, R.J.D. Menezes, J. De Klerk, I.R. Castleden, C.M. Hooper, G. Carneiro, M. Gilliham, «Identifying protein subcellular localisation in scientific literature using bidirectional deep recurrent neural network», Scientific Reports, 11(1), 1-11, 2021, doi:10.1038/s41598-020-80441-8.

[49] B. Dreyfus, A. Chaudhary, P. Bhardwaj, V.K. Shree, «Application of natural language processing techniques to identify off-label drug usage from various online health communities», Journal of the American Medical Informatics Association, 28(10), 2147-2154, 2021, doi:10.1093/jamia/ocab124.

[50] M.J. Acosta, G. Castillo-Sánchez, B. Garcia-Zapirain, I. de la Torre Díez, M. Franco-Martín, «Sentiment analysis techniques applied to raw-text data from a csq-8 questionnaire about mindfulness in times of covid-19 to improve strategy generation», International Journal of Environmental Research and Public Health, 18(12), 2-21, 2021, doi:10.3390/ijerph18126408.

[51] P. Fairie, Z. Zhang, A.G. D'Souza, T. Walsh, H. Quan, M.J. Santana, «Categorising patient concerns using natural language processing techniques», BMJ health & care informatics, 28(1), 1-9, 2021, doi:10.1136/bmjhci-2020-100274.

[52] T. Basu, S. Goldsworthy, G. V. Gkoutos, «A sentence classification framework to identify geometric errors in radiation therapy from relevant literature», Information (Switzerland), 12(4), 1-11, 2021, doi:10.3390/info12040139.

[53] A. Borjali, M. Magnéli, D. Shin, H. Malchau, O.K. Muratoglu, K.M. Varadarajan, «Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation», Computers in Biology and Medicine, 129, 1-26, 2021, doi:10.1016/j.compbiomed.2020.104140.

[54] S.K. Prabhakar, D.-O. Won, «Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention», Computational Intelligence and Neuroscience, 2021, 1-16, 2021, doi:10.1155/2021/9425655.

[55] N. Ali, A.H. Abuel-Atta, H.H. Zayed, «Enhancing the performance of cancer text classification model based on cancer hallmarks», IAES International Journal of Artificial Intelligence, 10(2), 316-323, 2021, doi:10.11591/ijai.v10.i2.pp316-323.

[56] N. Khamphakdee, P. Seresangtakul, «Sentiment analysis for Thai language in hotel domain using machine learning algorithms», Acta Informatica Pragensia, 10(2), 155-171, 2021, doi:10.18267/j.aip.155.

[57] S. Madichetty, M. Sridevi, «A stacked convolutional neural network for detecting the resource tweets during a disaster», Multimedia Tools and Applications, 80(3), 3927-3949, 2021, doi:10.1007/s11042-020-09873-8.

[58] N. Cerkez, B. Vrdoljak, S. Skansi, «A Method for MBTI Classification Based on Impact of Class Components», IEEE Access, 20(2017), 1-19, 2021, doi:10.1109/ACCESS.2021.3121137.

[59] P. Kulkarni, K.N. Cauvery, «Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique», International Journal of Advanced Computer Science and Applications, 12(9), 508-517, 2021, doi:10.14569/IJACSA.2021.0120957.

[60] Y. Li, P. Xu, Q. Ruan, W. Xu, «Text Adversarial Examples Generation and Defense Based on Reinforcement Learning», Tehnicki vjesnik - Technical Gazette, 28(4), 1306-1314, 2021, doi:10.17559/TV-20200801053744.

[61] M. Zulqarnain, R. Ghazali, M.G. Ghouse, N.A. Husaini, A.K.Z. Alsaedi, W. Sharif, «A comparative analysis on question classification task based on deep learning approaches», PeerJ Computer Science, 7, 1-27, 2021, doi:10.7717/PEERJ-CS.570.

[62] A. Moreo, A. Esuli, F. Sebastiani, «Word-class embeddings for multiclass text classification», Data Mining and Knowledge Discovery, 1-29, 2021, doi:10.1007/s10618-020-00735-3.

[63] W.H. Park, N.M.F. Qureshi, D.R. Shin, «Pseudo nlp joint spam classification technique for big data cluster», Computers, Materials and Continua, 71(1), 517-535, 2022, doi:10.32604/cmc.2022.021421.

[64] D.T. Tolciu, C. Săcărea, C. Matei, «Analysis of patterns and similarities in service tickets using natural language processing», Journal of Communications Software and Systems, 17(1), 29-35, 2021, doi:10.24138/JCOMSS.V17I1.1024.

[65] S.K. Prabhakar, H. Rajaguru, D.O. Won, «Performance Analysis of Hybrid Deep Learning Models with Attention Mechanism Positioning and Focal Loss for Text Classification», Scientific Programming, 2021, 1-12, 2021, doi:10.1155/2021/2420254.

[66] H. Ali, M.S. Khan, A. AlGhadhban, M. Alazmi, A. Alzamil, K. Al-utaibi, J. Qadir, «All Your Fake Detector Are Belong to Us: Evaluating Adversarial Robustness of Fake-news Detectors Under Black-Box Settings», IEEE Access, 4(2016), 1-15, 2021, doi:10.1109/ACCESS.2021.3085875.

[67] L. Shi, Y. Zhu, Y. Zhang, Z. Su, «Fault Diagnosis of Signal Equipment on the Lanzhou-Xinjiang High-Speed Railway Using Machine Learning for Natural Language Processing», Complexity, 2021, 1-13, 2021, doi:10.1155/2021/9126745.

[68] L. Burdick, J.K. Kummerfeld, R. Mihalcea, «To batch or not to batch? Comparing batching and curriculum learning strategies across tasks and datasets», Mathematics, 9(18), 2021, doi:10.3390/math9182234.

[69] E. Negro-Calduch, N. Azzopardi-Muscat, R.S. Krishnamurthy, D. Novillo-Ortiz, «Technological progress in electronic health record system optimization: Systematic review of systematic literature reviews», International Journal of Medical Informatics, 125, 1-8, 2021, doi:10.1016/j.ijmedinf.2021.104507.

APPENDIX A. NLP TECHNIQUES

Dimension	**Definition**
Word segmentation (Tokenize)	It is the process of converting paragraphs into inputs for the computer through a word list.
Data cleanup (Stop word)	It is the process of removing words that do not add exclusionary meaning to a sentence.
Lexicographic analysis with stemming	It is the process of converting each word of the sentence to its root form by removing or replacing suffixes.
Lexicographic analysis with lemmatization	It is a more accurate process than stemming and involves making an analysis of the vocabulary and its morphology to return to the basic form of the word.

APPENDIX B. NLP ALGORITHMS

Type	**Algorithm**	**Definition**
Basic mathematical functions	POS	It is the process of grammatical tagging or disambiguation of word categories.
	Name entity recognition (NER)	It is the process of "finding out" if a piece of data belongs to a person or business organization.
	N-gram	It is a sub-sequence of n items of a text data sequence. It is a probabilistic algorithm that allows making a statistical prediction of the next item of a sequence of a string of text data.
	Bag of words (BoW)	It is the process that allows the feature extraction from the text, determines the number of times that there is a word in the sentence.
Basic statistical algorithms	Term frequency - inverse document frequency (TF-IDF)	It is a statistical model that allows scoring the data to reflect their relevance in a given document.
	Text vectorization	It is the process which transforms the input of language into something that the computer can understand
	Statistical standardization	It is the process used to scale features of document data.

APPENDIX C. PLN ALGORITHMS OF THE 46 ARTICLES.

Author	**PLN Algorithms**	**ML Algorithms**
[17]	BoW, TF-IDF, N-Gram	SVM, NB, K-Medias
[63]	TF-IDF	SVD, R
[43]	TF-IDF	R, SVM, ANN
[44]	N-gram, TF-IDF	RF, KNN, AD
[48]	CBoW	LSTM, RNN
[9]	TF-IDF	R, LSTM, NB, RF, SVM
[30]	TF-IDF	NB, RF, AD
[16]	BoW, TF-IDF	ANN, RNN, LSTM
[49]	TF-IDF	SVM, NB, RF, RNN
[68]	Word2Vect	LSTM

[32]	CBoW, TF-IDF	CNN, LSTM
[42]	TF-IDF	LSTM
[33]	Glove	LSTM
[60]	Word2Vect	NB, SVM, MLP, CNN, RNN, LSTM, GRU
[34]	N-Gram	R, NB, RF
[50]	TF-IDF, Glove, Skip Gram	MLP
[51]	BoW	LDA
[52]	BoW, Skip Gram	SVM
[28]	Word2vect, Glove	RNN, LSTM
[29]	N gram	MLP
[64]	BoW, Word2vec, Glove	LSTM
[24]	TF-IDF, Word2Vec	SVM
[53]	N gram	KNN, RF, SVM, CNN, LSTM
[35]	BoW	NB
[36]	TF-IDF	NB, MLP, SVM, KNN
[59]	BoW, Word2Vect	LSTM
[65]	Word2Vect	RNN, LSTM
[58]	TF-IDF, BoW, Glove	NB, KNN, SVM, LSTM, R
[41]	Glove	CNN
[37]	TF-IDF, Word2Vec	LSTM, CNN
[56]	Word2Vec, BoW, TF-IDF	SVM, NB, R, RF, AD, PA, ADA
[54]	Word2Vec, Glove	LSTM, GRU
[38]	Glove	KNN, NB, PA, LSTM
[61]	Word2Vec	LSTM, GRU, CNN
[67]	BoW	LDA, SVM, NB, R, RF, KNN
[39]	TF-IDF	NB, R, AD, RF
[40]	Glove	LSTM
[66]	Glove	MLP, CNN, RNN
[25]	TF-IDF, N Gram, BoW	LDA, GRU, CNN
[55]	Glove	SVM, RNN, CNN
[45]	Word2Vec	CNN, RNN, LSTM
[62]	TF-IDF, Glove	SVM, CNN, LSTM
[46]	TF-IDF, Glove	NB, R, SVM, LSTM
[47]	TF-IDF, Word2Vec, Glove	SVM, KNN, NB, RF, AD, LSTM
[27]	TF-IDF, Word2Vec	LSTM
[57]	Word2Vec	SVM, KNN, CNN

APPENDIX D. MACHINE LEARNING ALGORITHMS – ML

ML Type	**Algorithm**	**Overview**
Supervised learning	K-NN	It is an algorithm that can be used to classify new samples or to predict values by looking for the "most similar" data points (by proximity).
	Regression - R	It is the algorithm that determines the relationships between dependent and independent variables for prediction and prognosis.
	Decision tree - DT	It is an algorithm that uses the fork for every possible outcome of a decision.
	Support Vector Machines - SVM	It is an algorithm that seeks to find a hyperplane that best separates two different kinds of data points.

	Naive Bayes - NB	It is a classification algorithm based on Bayes' theorem and classifies each value as independent from any other. It uses probability to predict a class or category.
	Radom Forest - RF	It is the algorithm that represents a set of decision trees in which each tree trains with different data samples from the same problem.
	Passive aggressive – PA	It is an algorithm that is used for large-scale learning. Input data come in sequential order and the machine learning model is updated step by step.
	Singular Value Decomposition – SVD	It is the algorithm used to eliminate redundant data. It determines which values are important and removes those that are not.
	AdaBoost – ADA	It is an algorithm that can be used together with other learning algorithms to improve its performance.
Unsupervised learning	Cluster K-Media	It is an algorithm that trains and properly knows the data to find hidden groups.
	Latent Dirichlet Allocation – LDA	It is an algorithm and its objective is to find the topics to which a document belongs based on the words it contains.
Deep Learning	Convolutional neural networks -CNN	It is one of the variants of neural networks that is widely used in the computer vision field.
	Recurrent Neural Networks – RNN	It is an algorithm widely used in natural language processing. It is used to analyze time series data.
	Long Short Term Memory - LSTM	It is an algorithm that introduces loops in the network diagram to memorize previous states of variables to decide which one will be next.
	Artificial neural network - ANN	It is a group of multiple perceptrons/neurons in each layer.
	Multilayer Perceptron - MLP	It is a network class consisting of at least three layers of nodes: an input layer, a hidden layer, and an output layer.
	Gating Circulation Unit – GRU	It is an enhanced RNN.