Journal of Information Systems and Informatics
Vol. 5, No. 3, September 2023 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v5i3.549 Published By DRPM-UBD

Bibliometric Analysis of Deep Learning for Social Media Hate Speech Detection

Raymond Mutanga[1], Oludayo Olugbara[2], Nalindren Naicker[3]

[1,2, 3]MICTSETA 4IR Center of Excellence, Durban University of Technology, South Africa
[1]raymutanga@gmail.com, [2]oludayoo@dut.ac.za, [3]naickern@dut.ac.za

Abstract

Social media has become an important web technology for creating and sharing information plus enhancing business reputations worldwide. However, the anonymity accorded by social media platforms has been cryptically vituperated to spread horrendous content such as hate speech. Recently, researchers have been progressively gravitating towards the use of deep learning techniques to address the problem of social media hate speech detection. This study provides bibliometric analysis and mapping of the existing literature on hate speech detection using deep learning algorithms. The study used articles published between 2016 and 2022 from the Scopus database, while Vos Viewer, Biblioshiny, and Panda's software tools were employed for the bibliometric analysis. The research explored the yearly trajectory of recent publications, dominant countries, collaborative institutions, sources of primary studies that have employed deep learning for hate speech detection, and the intellectual and social structures of the research constituents. It has been observed that the literature on hate speech detection is rapidly growing, but research output and collaborations from the developing countries of the world are still limited. The findings of this study provide insights into the intellectual structure and advancements in deep learning applications for hate speech detection while identifying research gaps for future work.

Keywords: Bibliometric, Deep Learning, Hate Speech

1. INTRODUCTION

The advent of Web 2.0 technologies such as Twitter and Facebook, has completely revolutionised information communication by allowing users in disparate geographical locations of the world to use social media platforms to seamlessly create, discover, congregate, share, communicate, and exchange information with people. The platforms facilitate the effective integration of people of different cultures, heritages, and religions to interact and build relationships and business reputations across the world. Large volumes of user-generated content are continuously produced and posted on social media platforms daily. For instance, In 2017, Twitter had 330 million active users per month, and 157 million of the users were active daily, sharing approximately 500 million tweets each day [1].

Given this meteoric rise of user-generated content on social media platforms such as Twitter, the volume of online hate speech started growing exponentially [2]. In response to this trend, social media organizations have formulated internal regulatory policies about hate speech proliferation on their platforms and became signatories to the European Commission code of conduct [3].

Machine learning-based hate speech recognition methods have been proposed in response to the shortcomings of human annotators and legislation. Classical and deep learning algorithms are the two taxonomic subclasses of machine learning methods for solving hate speech problems on social media platforms. Classical learning algorithms make use of handcrafted features, such as simple surface features, word generalisation features, lexical resources, and meta-information [2]. Several hate speech detection studies have used simple surface features such as Bag of Words and n-grams as input put to classical algorithms such as Support Vector Machine and Naïve Bayes [4-6]. The Bag of Words approach has been criticised for having a high positive rate, consequently, other studies have explored other sophisticated methods to generate salient features for classical machine learning methods [4, 7, 8]. However, the use of manually engineered features is time-consuming and the features are ordinarily insufficient to adequately address the problem of hate speech propagation on social media platforms [9]. In particular, manually engineered features fail to effectively capture the semantic and domain-specific representations of text documents[10-12]. Furthermore, individual classical algorithms have been criticised for their susceptibility to high variance, thereby negatively impacting predictive efficacy [13].

To address the aforementioned challenges associated with individual classical algorithms, other scholars have investigated the technique of ensemble learning for hate speech detection. Mutanga, et al. [13] combined Logistic Regression, Decision Trees, and Support Vector Machines using Voting to detect hate speech on Twitter. Their proposed ensemble approach outperformed Individual algorithms trained on the same dataset. Ahluwalia et al. in [14] employed an ensemble learning approach that integrated Logistic Regression, Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machine (GBM) algorithms to detect instances of hate speech directed towards women. The model was trained on both binary and multiclass datasets and achieved an optimal accuracy rate of 65.10% for binary classification and an F1-score of 40.60% for multiclass classification. Aljero and Dimililer [15] proposed a stacked ensemble for detecting hate speech using Logistic Regression, XGBoost, and Support Vector Machine as classifiers, and word2vec features. Their approach outperformed individual base classifiers on three different datasets. Despite the superior performance of ensemble learning approaches for hate speech detection, it is worth noting that most ensemble learning approaches are based on classical base algorithms, therefore we argue that they are also susceptible to problems associated with manual feature engineering.

Research in automated hate speech detection has gravitated towards deep learning algorithms, that carry out end-to-end training with huge datasets, allowing the encoding of salient feature representations. A recent analysis of datasets and classifiers revealed that hate speech lacks unique, discriminative features and therefore the task may be challenging for models that depend on manually engineered features[16]. Deep learning algorithms can capture complex data representations which makes them applicable for identifying hate speech, where the natural human language used is highly ambiguous in word senses. For instance, the problem of ungrammatical text rife in Twitter data has largely been mitigated by deep neural architectures [17].

The shift in focus from classical machine learning to deep learning has prompted this study, which employs bibliometric analysis to systematically explore existing research on hate speech detection using deep learning methods. Bibliometric analysis is a field of study that attempts to use bibliographic data extracted from past publications and their citation relations to evaluate and reveal the structure of research. Previous research [18-21] used bibliometric methods to analyse different subfields of computer science. The primary objective of this study is to unveil the structure and dynamics of hate speech detection studies, guided by the following four research questions:

(a) What is the relative importance of articles from journals and conferences in addressing the problem of social media hate speech detection?
(b) What is the research productivity of institutions and countries in the applications of deep learning methods for social media hate speech detection?
(c) What is the intellectual and social structure of the previous authors that have researched deep learning for social media hate speech detection?
(d) What are the important concepts in the automation of social media hate speech detection using deep learning methods?

Through this investigation, we aim to provide a comprehensive understanding of the evolution and current landscape of deep learning-based hate speech detection literature, offering insights into the prominence of different publication sources, regional contributions, author networks, and pivotal concepts that underpin this critical domain. The rest of this article is organised as follows. Section 2 introduces the study methodology and provides the details of the selected electronic database. Section 3 presents the study results with a special focus on performance analysis and science mapping. Section 4 discusses the results of the study, and Section 5 gives a conclusive remark, including the study's limitations and future work.

2. METHODS

The electronic database of Scopus was selected as the main source of data for this study. Scopus is a famous peer-reviewed database of research papers in areas such as Science and Technology, Engineering, Humanities, Social sciences, and Health

sciences. Results from a recent study have revealed that Scopus is one of the most dependable databases and has a high level of consistency on author names, volume, and issue numbers [22]. Furthermore, a comparison of the Web of Science and Scopus revealed that Scopus provides a wider field coverage in natural sciences and engineering disciplines [23]. The scholastic database provides a sufficient volume of articles on hate speech detection using deep learning methods. The subsequent section outlines the search query used to extract articles from the Scopus dataset, following the preferred reporting items for systematic reviews and meta-analysis (PRISMA) as a guiding protocol [24, 25].

PRISMA is a widely used and internationally recognised reporting guideline for systematic reviews and meta-analyses. PRISMA provides a transparent and standardised framework for researchers to report their methods and findings, which allows readers to evaluate the quality and rigor of the study. Figure 1 outlines the sequential steps followed in selecting articles included in this study using PRISMA.

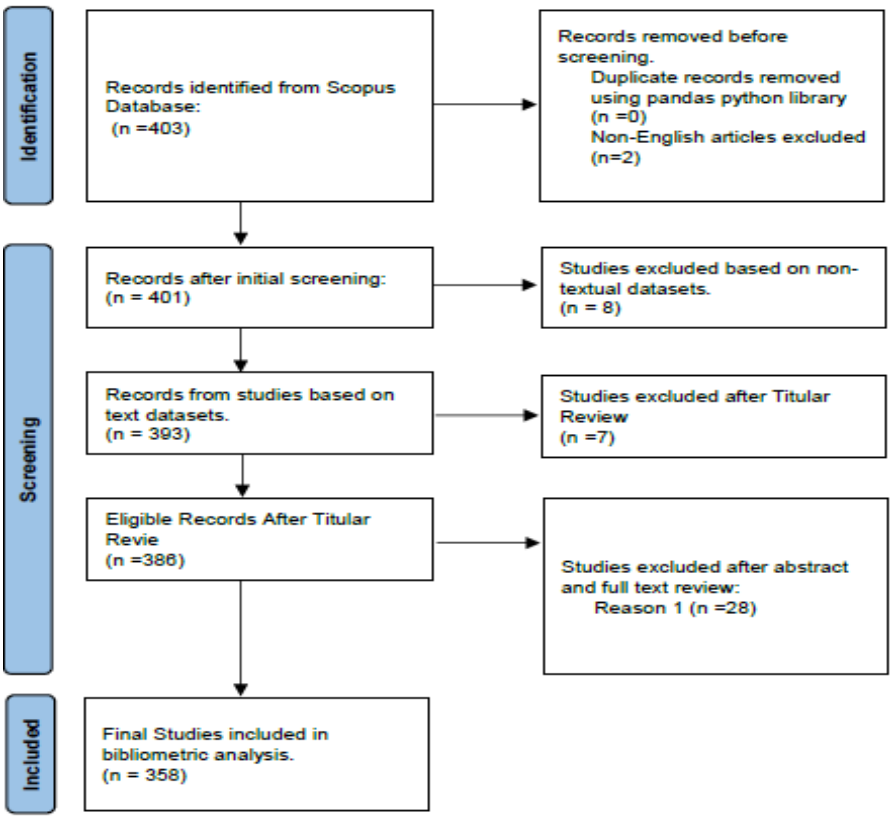

Figure 1. Flow diagram illustrating the systematic selection process of studies included in the meta-analysis following PRISMA Guidelines

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

The internet search for related articles to congregate data was implemented using keywords that are relevant to the theme of the present work. The article inclusion and exclusion criteria are articles that were published in the English language between the years 2016 and 2022. Moreover, keywords selected for searching mechanisms should be relevant to the theme of the present study and as exhaustive as possible. The keywords used in this study are "Hate Speech", "Offensive Speech", "Deep Learning" and "Deep Neural Network". The syntax for the search string that was used in the present study to discover the relevant articles is as follows. (TITLE-ABS-KEY ("hate speech" OR "Offensive language") AND TITLE-ABS-KEY (detect* OR recogn*) AND TITLE-ABS-KEY ("deep learning" OR "deep neural network")) AND PUBYEAR > 2016 AND PUBYEAR < 2023 AND PUBYEAR > 2016 AND PUBYEAR < 2023. This search string extracts the relevant articles containing the stated keywords in their title, abstract, and published between the years 2016 and 2022. Studies that were published in restricted languages that are not English were excluded from this study to avoid the unnecessary costs of translation.

Data congregated from the discovered articles were further cleaned to remove duplicate records and records with missing values after extracting the file. Data cleaning is an essential step of bibliometric analysis because public databases such as Scopus are not essentially designed for bibliometric analysis. The removal of duplicates, missing values, and other erroneous records on the downloaded comma-separated value (csv) file was achieved programmatically using the Pandas library embedded in Python programming language. The final database for analysis contained 358 articles after the screening and data-cleaning processes. The VosViewer and Biblioshiny library from the R studio were used to perform the analysis of the extracted data [26]. The experiments conducted encompass performance analysis and bibliographic coupling to assist in locating highly related previous studies.

3. RESULTS

The section on results is alienated into performance-based analysis and science mapping-based analysis as subsequently explicated.

3.1 Performance-based Analysis.

Performance-based analysis in bibliometrics study examines the contributions of research constituents such as authors, institutions, and countries of a field [27, 28]. This section presents the results of a performance-based analysis for hate speech detection using deep learning methods. The results are based on the yearly trajectory of hate speech detection using deep learning, dominant sources of articles, countries where the research was done, and document citations.

Figure 2 illustrates the publication trajectory of hate speech detection using deep learning methods between 2016 and 2022. Since 2016, it can be observed that the number of publications has been increasing. This growth in publications on hate speech detection using deep learning can be attributed to the increased availability of huge datasets that allow researchers to apply deep learning methods which require a lot of data for model training and evaluation. In addition, the advent of enhanced deep learning methods such as the transformer in 2017 [29] can be attributed to the sharp rise in the number of publications, notably between 2018 and 2019. The transformer-based deep learning methods were developed to address the inherent difficulty such as disambiguation in processing natural languages. Figure 2 illustrates the volume of yearly publications on hate speech detection using deep learning methods. It can be observed in the figure that between 2016 and 2018, there was slow progress in the application of deep learning methods for hate speech detection. However, there was rapid progress between 2018 and 2021, after which a slight decline in progression was observed in 2022.

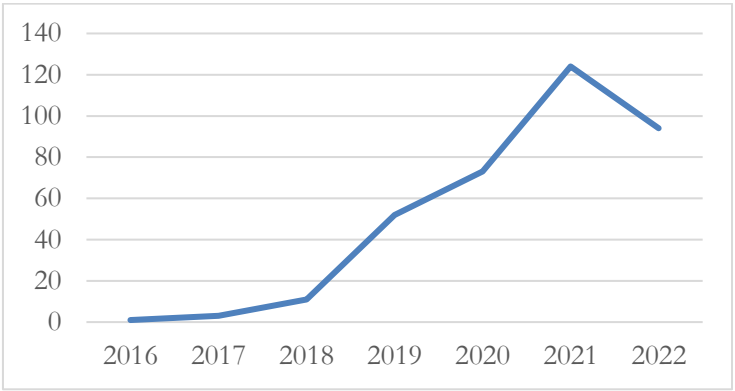

Figure 2. Yearly publications on hate speech detection using deep learning methods

Table 1 outlines the top ten dominant sources of conference and journal articles that published research works on hate speech detection using deep learning methods. In addition, the number of articles published by each source is indicated. The dominant source of information on hate speech detection using deep learning methods was reported by CEUR workshop proceedings which published 52 articles during the period considered. The lecture notes in computer science also contributed a significant 26 articles followed by IEEE Access with 11 publications, while the rest of the sources contributed less than 11 articles to the total. It is worth noting that the majority of the top ten article sources are conference proceedings. This trend can be attributed to the fact that journals are more stringent in the peer review and screening process for publication. Therefore, fewer researchers successfully get their articles published by top-quality journals.

The delay in getting research articles to be published by most top journals is also contributing to most authors patronising conferences.

Table 1. Top ten sources of hate speech detection using deep learning methods.

Sources	Articles
CEUR Workshop Proceedings	52
Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	26
IEEE Access	11
Acm International Conference Proceeding Series	10
14th International Workshops on Semantic Evaluation, Semeval 2020 - Co-Located 28th International Conference on Computational Linguistics, Coling 2020, Proceedings	9
NAACL HLT 2019 - International Workshop on Semantic Evaluation, Semeval 2019, Proceedings of the 13th Workshop	8
Applied Sciences (Switzerland)	6
Communications in Computer and Information Science	6
Social Network Analysis and Mining	6
Advances in Intelligent Systems and Computing	5

Table 2 presents the result of the impact analysis of hate speech detection publication sources based on the metrics of h-index and total citations (TC). The h-index is an objective measure of research productivity that combines both citations and publications to measure the performance of a research constituent [27]. The measure indicates the quality and consistency of a research constituent over the recent period of five years [30]. The number of citations is the collective times that articles from a given publication source have been cited. It can be noted in Table 2 that the most impactful source of articles was CEUR workshop proceedings with an h-index of 7. Although Lecture Notes in Computer Science has a total citation of 352, CEUR workshop proceedings with 178 citations have a stronger h-index. This result indicates that more articles from CEUR proceedings have received citations of at least 7 as compared to research papers from the Lecture Notes in Computer Science. In addition, the ACM Proceeding series, IEEE Access RANLP had relatively impactful research output during the period considered.

Table 2. Top ten impactful publication sources

Sources	h-index	TC
CEUR Workshop Proceedings	7	178
Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	6	352

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Sources	h-index	TC
ACM International Conference Proceeding Series	5	245
IEEE Access	5	162
International Conference Recent Advances in Natural Language Processing, RANLP	4	217
14th International Workshops on Semantic Evaluation, Semeval 2020 - Co-Located 28th International Conference on Computational Linguistics, Coling 2020, Proceedings	3	21
Electronics (Switzerland)	3	28
Expert Systems with Applications	3	61
Information Processing and Management	3	75
NAACL HLT 2019 - International Workshop on Semantic Evaluation, Semeval 2019, Proceedings of the 13th Workshop	3	29

Figure 3 shows the result of the most productive countries by volume of publications that have contributed to good research on hate speech detection using deep learning methods. Most articles can be observed to be authored by researchers from India (308 articles, 85%). The remaining countries have contributed fewer articles than the highest-ranked India, wherein the United States of America recorded 77 articles (22%), followed by Spain (67 articles, 19%), and Germany (40 articles,11%). The other countries that have contributed to the theme of hate speech detection using deep learning techniques include the United Kingdom, Pakistan, Saudi Arabia, Italy, China, and Indonesia. It is worth noting that the sum of the percentages surpasses 100% because a given article may have co-authors from different countries. Moreover, it is surprising to discover that no African country has featured in the top productive countries where hate speech detection research was reported in the outlets discussed.

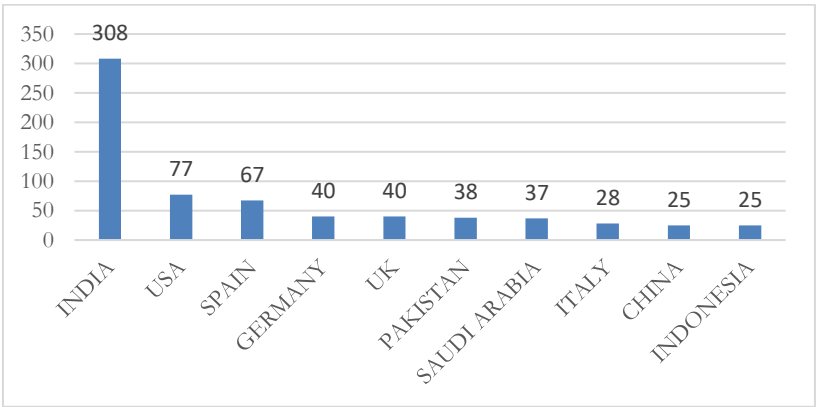

Figure 3. Top ten countries by volume of publication output

Table 3 outlines ten countries with the most cited articles on hate speech detection using deep learning methods. The research impact is measured based on total citations (TC) and the average of citations per article (ACA). The ACA, according to Clarivate is the sum of the times cited count divided by the number of results discovered. Saudi Arabia has the highest number of total citations (161), followed by the United Kingdom (136), followed by India (104), and Norway (100). However, it is worth noting, that although India has a higher research output by volume as depicted in Figure 3, its total citations are lower than those of Saudi Arabia and the United Kingdom. This suggests that research output from Saudi Arabia and the United Kingdom could be more visible, and applicable when compared to those from India. The lesser applicability of the research outputs from India is further demonstrated by the fact that it also recorded the lowest average citations as depicted in Table 3. Figure 3 illustrates that there is no predominant African country in terms of voluminous production. Nevertheless, Table 3 reveals that Egypt holds the distinction of being the sole African country with the highest impact in terms of research output.

Table 3. Top ten countries with the most impactful publications

Country	TC	ACA
Norway	100	100.00
United Kingdom	136	34.00
Germany	71	23.67
Egypt	91	22.75
Saudi Arabia	161	20.13
China	68	11.33
Pakistan	45	11.25
France	38	9.50
Spain	41	5.86
India	104	3.25

Table 4 lists the results of the most cited articles on hate speech detection using deep learning published between 2016 and 2022. It can be observed that the article on hate speech detection in tweets is the most cited with 552 citations [31]. The main author of the paper is Pinkesh Badjatiya from the IIIT-H in Hyderabad India. The article that explored the use of convolution-GRU-based deep neural networks to detect hateful tweets was ranked second with 290 citations. Moreover, it can be observed that five articles on the list had less than 100 citations. The total citations (TC) per year (TCY) gives a more objective assessment of the article quality because older articles naturally tend to get more citations by having been in existence for a longer period. An observation can be made that despite Article 4 receiving more citations than Article 5, its TCY value is comparatively lower.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Table 4. Performance analysis of articles on hate speech detection

No.	Article DOI	TC	TCY
1	10.1145/3041021.3054223	552	92.00
2	10.1007/978-3-319-93417-4_48	290	58.00
3	10.1145/3368567.3368584	128	32.00
4	10.26615/978-954-452-049-6-062	117	19.50
5	10.1007/s10489-018-1242-y	100	20.00
6	10.1109/ASONAM.2018.8508247	81	16.20
7	10.3233/SW-180338	76	19.00
8	10.1109/ACCESS.2019.2899260	76	19.00
9	10.26615/978-954-452-049-6-036	75	12.50
10	10.1080/03772063.2022.2043786	72	12.70

3.2 Science Mapping-Based Analysis

The denotation of science mapping is to examine linkages among the given research constituents. Results from a science mapping experiment reveal the association and structural connections among research constituents. The science mapping techniques explored in this study include co-authorship analysis, bibliographic coupling, and co-word analysis as expounded in this section.

Figures 4 and 5 respectively visualise the scientific co-authorship relationships among the most collaborative universities and countries across the world. The size of the affinity network nodes has depicted the number of joint publications whereas the strength of the collaborations is indicated by the thickness of the lines connecting the network nodes. In the co-authorship analysis, the minimum number of articles per institution and the minimum number of articles per country were respectively set at 3. Four connections depicted by different colors have emerged from the co-authorship analysis. Da-LICT, a private University in Gandhinagar from India was the most collaborative Institution with a total link strength of 6. The Eastern University in Chenkalady Sri Lanka was the second most collaborative institution with a total link strength of 3. In addition, the LDRP Institute of Technology and Research in Gandhinagar India had significant collaborations with other institutions.

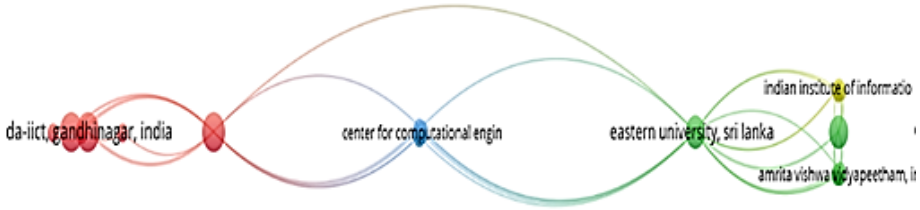

Figure 4. Most collaborative universities on social media hate speech detection research

Figure 5 illustrates the most collaborative network of countries where researchers are studying social media hate speech detection using deep learning methods. The country collaborative network reports a minimal publication count of three as illustrated in Figure 5. India, the United Kingdom, and Germany were the most collaborative countries where researchers are studying social media hate speech detection using deep learning methods. There is strong collaboration between India and countries such as Sri Lanka, Brazil, and Ireland. The United Kingdom also has strong research collaborations with France, China, and Finland. Moreover, it can be observed that collaborative work from African countries is highly limited. Furthermore, it can be seen in Figure 5 that among African countries, only Morocco and Egypt have collaborated with more than one country.

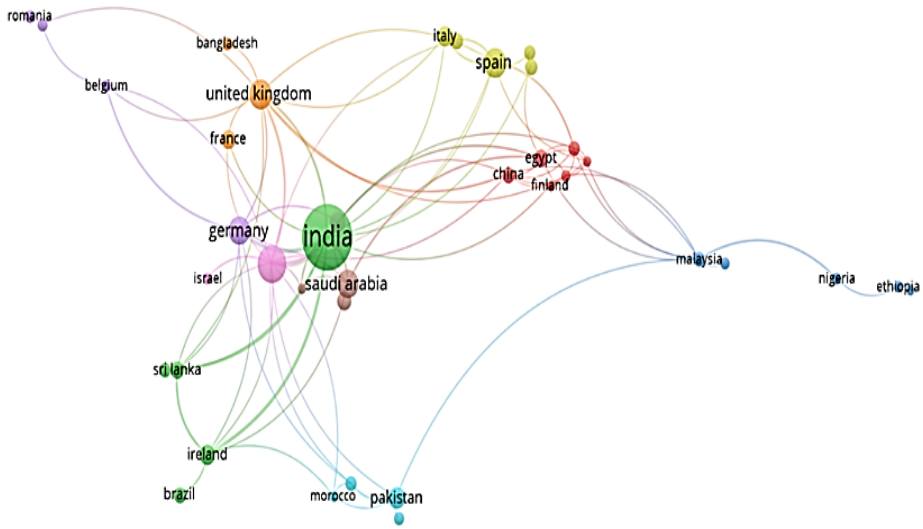

Figure 5. Research collaboration network of countries

The concept of bibliographic coupling assumes that two publications sharing common references are also similar. It creates thematic clusters of research articles based on shared preferences using a clustering method. The VosViewer software employs mapping and clustering methods to objectively evaluate affinity strength

where sources are clustered using unique colors to identify them [32, 33]. In addition, a relevant distance-based map is plotted to highlight clusters of article authors or journals where the articles were published [34]. The clustering provides an overview to the researchers for reading an article for submitting their manuscript to the relevant journal. The high instances of shared references indicate similar intellectual capital [35]. Figure 6 illustrates the bibliographic coupling of the top 26 authors who have studied social media hate speech detection using deep learning methods. Four major thematic clusters and two smaller clusters have emerged from the bibliographic coupling of authors. The most dominant thematic cluster depicted in red has Budi, Fohr, Nayak, Joshi, Chakraboryt, Rosso and Mishra, [36-44] [43, 45-49]. The second most dominant thematic cluster, Bhattacharya, Mirnalinee, Thenmozhi, Chakravarthi, and McCrae as authors [50, 51] [52-59]. The coupled authors signify the existence of common intellectual capital amongst the authors. Chavakarti was found to be the most influential author with a link strength of 846. In addition, Martin Valdivia, Del arco, and Urena Lopez[60-62] were found to have a significant influence in the study of social media hate speech detection using deep learning methods.

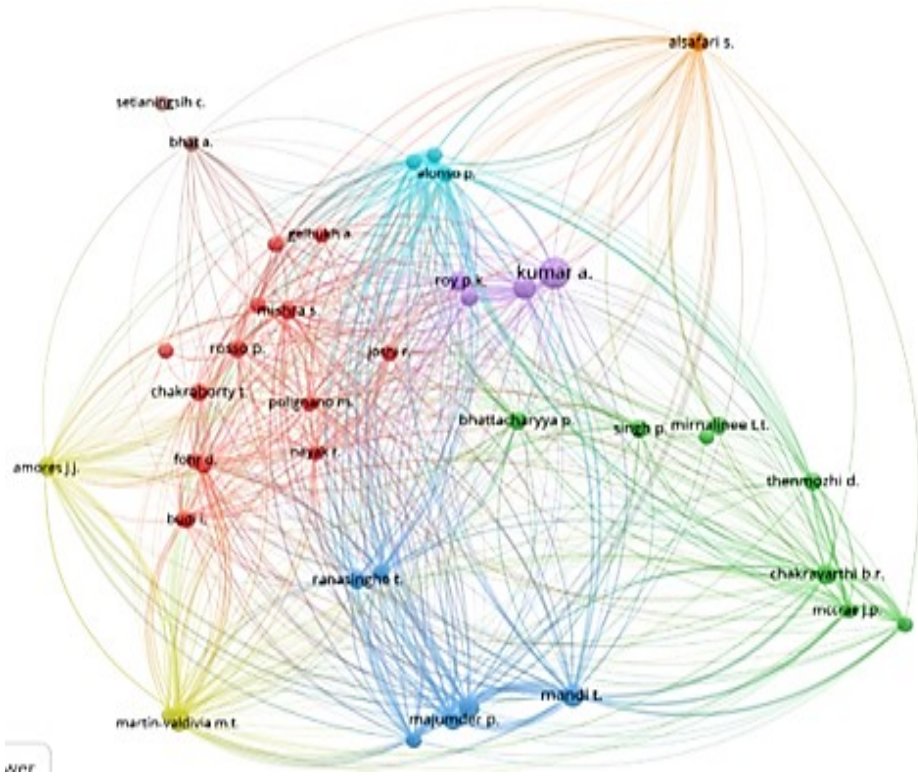

Figure 6. Bibliographic coupling of authors

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Figure 7 illustrates the bibliographic coupling of publication sources that have published articles on social media hate speech detection using deep learning methods. In bibliographic coupling experiments, the minimum number of articles per author and the minimum number of articles per source was set at 3. The CEUR workshop proceedings have exhibited the highest influence on deep learning-based hate speech detection literature which is closely followed by the Lecture Notes in Computer Science. Although bibliographic coupling is widely used for bibliometric analysis, certain researchers have criticised it for its inefficiency in analysing old publications [63]. However, it allows for the inclusion of publications that are not yet cited and is handy in predicting emerging and future trajectories [27, 63].

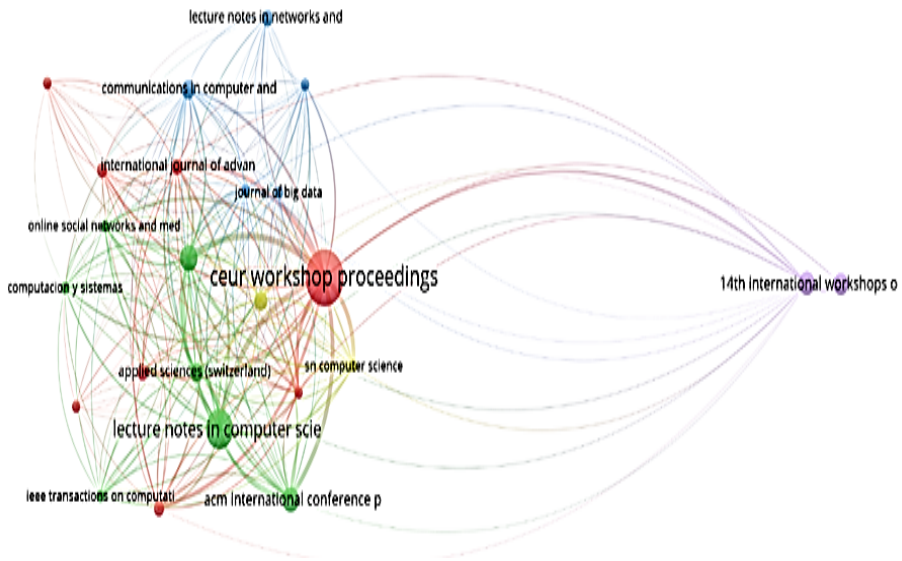

Figure 7. Bibliographic coupling of publication sources

Co-word analysis examines the content of research articles and keywords that summarise the literature on the area [64]. The scheme of co-word analysis assumes that words that frequently appear concomitantly, focus on similar themes. Keywords used in the analysis are extracted from the titles, abstracts, and full texts of articles [65, 66]. In this study, we examined co-word analysis for all keywords and author keywords. The minimum number of word co-occurrence used in the experiments was set at 10. Figure 8 categorises the relevant top co-occurring keywords from the publications used in this study into four clusters. Each line represents the relationship between two keywords such that two keywords are considered as being co-cited when they appear in tandem in the same article. The size of network nodes has reflected the frequency of keywords such that the higher the frequency of a keyword, the larger the size of a network node. The main keywords per cluster are deep learning 'hate speech' (green cluster), 'social media

(yellow cluster), and 'Twitter' (blue cluster). The red cluster illustrates the connections among 'deep learning,' 'nlp,' 'bert,' 'and transfer learning.' This cluster is logically explained by the fact that it depicts the state-of-the-art deep learning methods being used such as transformer algorithms [29] in the detection of hate speech in online spaces [67].

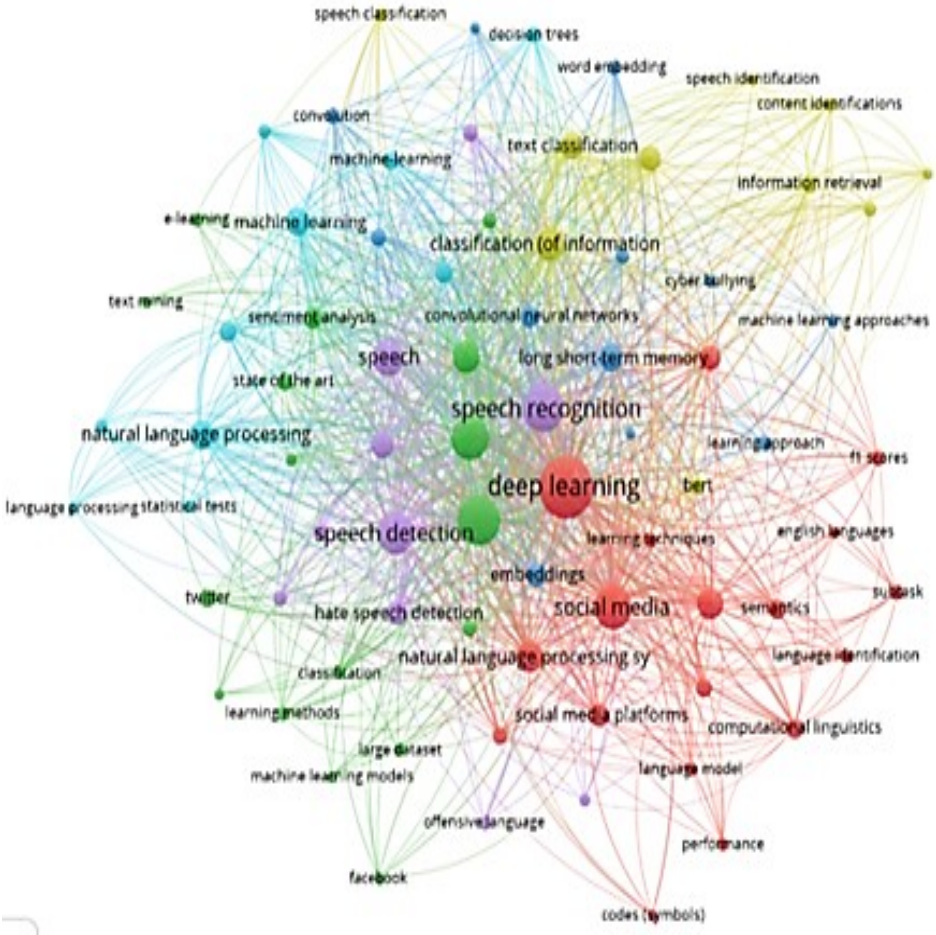

Figure 8 Co-word analysis of all keywords

Figure 9 graphically illustrates the top co-occurring author keywords from the publications used in this study. The green cluster shows the connections among the related concepts such as 'hate speech,' 'machine learning' and 'cnn'. These connections illustrate a stream of earlier forms of hate speech detection methods that were dominant before the advent of attention-enhanced models and transfer learning.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

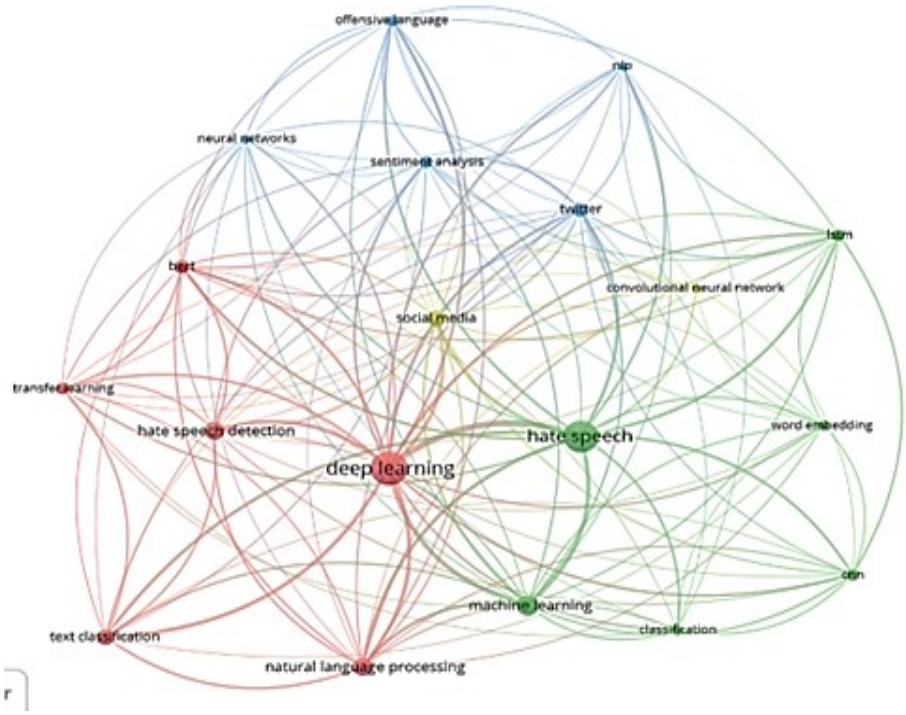

Figure 9. Co-word analysis of author keywords

3.2 Discussion

This study examined the progression of research articles on social media hate speech detection using deep learning methods. The study results confirm positive movements in the research endeavor with an upward trend in the number of yearly publications as shown in Figure 2. This trend is consistent with the growing interest and investment in deep learning as a transformative technique in Natural Language Processing tasks [9]. For instance, findings from a recent study pointed to a shift toward deep learning methods in hate speech detection[68]. In addition, the study has revealed a very low number of publications from African and South American countries. From Africa, only South Africa, Nigeria, Egypt, Morocco, and Ethiopia published articles on hate speech detection using deep learning techniques.

The analysis of article sources performed to address the first and second research questions of this study has shown that conferences are the favored publication approach by the researchers working in this Domain. This finding suggests that researchers and practitioners in this field value the timeliness and flexibility of conferences in sharing new research ideas and findings. Moreover, this trend may

also reflect the interdisciplinary nature of the research, which requires quick dissemination of ideas to keep up with the fast-changing landscape of social media. The dominant countries by publication volume in this domain are India, The United States, Germany, and the United Kingdom as depicted in Figure 3. However, Norway and the United Kingdom had more impactful publications as compared to the rest of the countries. Our analysis also revealed that research on social media hate speech detection using deep learning is concentrated in a few key regions notably Europe and the United States of America, and we argue limit the diversity of perspectives and approaches to the problem. This is particularly important since cultural variations affect the definition of what constitutes hate speech[68]. Science mapping experiments conducted in this study addressed the third research question of this study. The results from these experiments indicate the existence of collaborative work between different research institutions. Countries such as India and the United Kingdom are the most collaborative.

This study reveals extraordinarily little collaboration from and with African countries. From Africa Only Morocco, Egypt and Ethiopia had research collaborations with other countries The keyword analysis used to address the fourth research question of this study revealed a strong presence of words such as deep learning, and LSTM. Bert and Transfer learning as compared to words such as machine learning and support vector machine. These results suggest that more and more people are moving away from classical methods such as support vector machines and gravitating towards state-of-the-art deep learning techniques and transfer learning in addressing the problem of hate speech on social media. However, these approaches have been criticised for their lack of interpretability when making decisions[69]. The implementation of a manual appeal process may be deemed necessary for automated hate speech detection systems, as it holds practical significance. Our comprehensive investigation provides valuable insights into hate speech detection using deep learning methods and positions the findings within the global discourse on combating online hate speech. The findings of this study inform future research directions, guide policy considerations, and promote collaborative efforts in mitigating hate speech dissemination on social media.

4. CONCLUSION

This study investigated various aspects of research on social media hate speech detection using deep learning methods. The study utilised bibliometric analysis techniques to analyse a large Scopus dataset of articles published between 1 January 2016 and 31 December 2022. The research questions addressed in the study included the relative importance of articles from journals and conferences, research productivity of institutions and countries, the intellectual and social structure of previous authors, and important concepts in the automation of social media hate speech detection. Results from the study reveal that conferences are the dominant sources of publications for hate speech using deep learning. The

most productive regions in Hate speech detection are Europe and the United States, while research output from Africa is still limited.

The study identified a core group of influential authors who have contributed significantly to the field of deep learning for social media hate speech detection. These authors were found to be highly connected and influential in the network of researchers. Finally, the study found that important concepts in the automation of social media hate speech detection using deep learning methods included attention-based deep learning algorithms and transfer learning. The study provides valuable insights into the bibliometric analysis of research publications on social media hate speech detection using deep learning methods. The findings contribute to a better understanding of the relative importance of articles from journals and conferences, the research productivity of institutions and countries, the intellectual and social structure of previous authors, and important concepts in automation.

The results of the study can inform future research in the field and assist policymakers and practitioners in developing effective strategies for addressing hate speech on social media platforms. Additionally, the study's methodology and approach can be replicated in other research areas, providing a valuable contribution to the bibliometric analysis of research publications. We note that using our search string has resulted in the inclusion of papers that only use hate speech or offensive speech to motivate their work thereby excluding papers from closely related areas such as sarcasm detection. Nevertheless, we consider the number of articles from these related areas too insignificant to include in the study. Information such as author names and author Institutions are not standardised in the Scopus database. Manual correction of such information is time-consuming and often prone to errors. This implies that incorrect information may have been fed into the analysis software thereby distorting the results. It is, therefore, necessary to explore the advantages of other Software tools such as Scimat in the future.

Bibliometric analysis was limited to articles on the detection of textual hate speech only. However, hate speech on social media can be expressed in forms other than text, for example, pictures, audio, and video. Future work must include articles that include all forms of media found on social media. Most of the studies analysed in this work focused on the English Language, however social media users across the world use different languages to express their hatred. There is a need to develop datasets specific to under-resourced languages so that hate speech may be tackled effectively. In conclusion, this study successfully revealed the structure and dynamics of deep learning-based hate speech detection literature. These findings assist researchers to understand the state-of-the-art in automated hate speech detection thereby presenting ideas for future research in the area.

REFERENCES

[1] U. Kursuncu, M. Gaur, U. Lokala, K. Thirunarayan, A. Sheth, and I. B. Arpinar, "Predictive analysis on Twitter: Techniques and applications," in Emerging research challenges and opportunities in computational social network analysis and mining: Springer, 2019, pp. 67-104.

[2] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proceedings of the fifth international workshop on natural language processing for social media, 2017, pp. 1-10.

[3] N. Alkiviadou, "Hate speech on social media networks: towards a regulatory framework?," Information & Communications Technology Law, vol. 28, no. 1, pp. 19-35, 2019.

[4] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in Proceedings of the NAACL student research workshop, 2016, pp. 88-93.

[5] F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in Proceedings of the first Italian conference on cybersecurity (ITASEC17), 2017, pp. 86-95.

[6] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in Proceedings of the AAAI Conference on Artificial Intelligence, 2013, vol. 27, no. 1, pp. 1621-1622.

[7] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proceedings of the second workshop on language in social media, 2012, pp. 19-26.

[8] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proceedings of the international AAAI conference on web and social media, 2017, vol. 11, no. 1, pp. 512-515.

[9] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," ieee Computational intelligenCe magazine, vol. 13, no. 3, pp. 55-75, 2018.

[10] F. Alkomah and X. Ma, "A Literature Review of Textual Hate Speech Detection Methods and Datasets," Information, vol. 13, no. 6, p. 273, 2022.

[11] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: From text to predictions," Information, vol. 13, no. 2, p. 83, 2022.

[12] R. Mutanga, "A comparative study of deep learning algorithms for hate speech detection on Twitter," 2021.

[13] R. T. Mutanga, N. Naicker, and O. O. Olugbara, "Detecting Hate Speech on Twitter Network using Ensemble Machine Learning," International Journal of Advanced Computer Science and Applications, vol. 13, no. 3, 2022.

[14] R. Ahluwalia, H. Soni, E. Callow, A. Nascimento, and M. De Cock, "Detecting hate speech against women in english tweets," EVALITA Evaluation of NLP and Speech Tools for Italian, vol. 12, p. 194, 2018.
[15] M. K. A. Aljero and N. Dimililer, "A novel stacked ensemble for hate speech recognition," Applied Sciences, vol. 11, no. 24, p. 11684, 2021.
[16] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? the challenging case of long tail on twitter," Semantic Web, vol. 10, no. 5, pp. 925-945, 2019.
[17] G. Kovács, P. Alonso, and R. Saini, "Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources," SN Computer Science, vol. 2, pp. 1-15, 2021.
[18] A. Keramatfar and H. Amirkhani, "Bibliometrics of sentiment analysis literature," Journal of Information Science, vol. 45, no. 1, pp. 3-15, 2019.
[19] P. Sánchez-Núñez, M. J. Cobo, C. De Las Heras-Pedrosa, J. I. Peláez, and E. Herrera-Viedma, "Opinion mining, sentiment analysis and emotion understanding in advertising: a bibliometric analysis," IEEE Access, vol. 8, pp. 134563-134576, 2020.
[20] A. Sarirete, "A Bibliometric Analysis of COVID-19 Vaccines and Sentiment Analysis," Procedia Computer Science, vol. 194, pp. 280-287, 2021.
[21] H. Zhu and L. Lei, "The Research Trends of Text Classification Studies (2000–2020): A Bibliometric Analysis," SAGE Open, vol. 12, no. 2, p. 21582440221089963, 2022.
[22] L. S. Adriaanse and C. Rensleigh, "Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison," The Electronic Library, 2013.
[23] P. Mongeon and A. Paul-Hus, "The journal coverage of Web of Science and Scopus: a comparative analysis," Scientometrics, vol. 106, no. 1, pp. 213-228, 2016.
[24] C. T. Olugbara, M. Letseka, R. E. Ogunsakin, and O. O. Olugbara, "Meta-analysis of factors influencing student acceptance of massive open online courses for open distance learning," The African Journal of Information Systems, vol. 13, no. 3, p. 5, 2021.
[25] C. T. Olugbara, M. Letseka, and O. O. Olugbara, "A Systematic Review of Digital Storytelling as Educational Tool for Teaching and Learning in Southern Africa," Multimodal Learning Environments in Southern Africa, pp. 165-195, 2022.
[26] J. A. Moral-Muñoz, E. Herrera-Viedma, A. Santisteban-Espejo, and M. J. Cobo, "Software tools for conducting bibliometric analysis in science: An up-to-date review," Profesional de la Información, vol. 29, no. 1, 2020.
[27] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," Journal of Business Research, vol. 133, pp. 285-296, 2021.
[28] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "An approach for detecting, quantifying, and visualizing the evolution of a

research field: A practical application to the Fuzzy Sets Theory field," Journal of informetrics, vol. 5, no. 1, pp. 146-166, 2011.

[29] A. Vaswani et al., "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998-6008.

[30] V. Grech and D. E. Rizk, "Increasing importance of research metrics: Journal Impact Factor and h-index," vol. 29, ed: Springer, 2018, pp. 619-620.

[31] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759-760.

[32] R. Khodabandelou, N. Aleebrahim, A. Amoozegar, and G. Mehran, "Revisiting three decades of educational research in Iran: A bibliometric analysis," Iranian Journal of Comparative Education, vol. 2, no. 1, pp. 1-21, 2019.

[33] N. Van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," scientometrics, vol. 84, no. 2, pp. 523-538, 2010.

[34] A. Kalantari et al., "A bibliometric approach to tracking big data research trends," Journal of Big Data, vol. 4, no. 1, pp. 1-18, 2017.

[35] H. Shin and R. R. Perdue, "Self-Service Technology Research: A bibliometric co-citation visualization analysis," International Journal of Hospitality Management, vol. 80, pp. 101-112, 2019.

[36] E. Sazany and I. Budi, "Hate speech identification in text written in Indonesian with recurrent neural network," in 2019 International Conference on Advanced Computer Science and information Systems (ICACSIS), 2019: IEEE, pp. 211-216.

[37] M. O. Ibrohim, E. Sazany, and I. Budi, "Identify abusive and offensive language in indonesian twitter using deep learning approach," in Journal of Physics: Conference Series, 2019, vol. 1196, no. 1: IOP Publishing, p. 012041.

[38] E. Sazany and I. Budi, "Deep learning-based implementation of hate speech identification on texts in indonesian: Preliminary study," in 2018 International Conference on Applied Information Technology and Innovation (ICAITI), 2018: IEEE, pp. 114-117.

[39] A. G. d'Sa, I. Illina, D. Fohr, and A. Akbar, "Exploration of Multi-corpus Learning for Hate Speech Classification in Low Resource Scenarios," in Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings, 2022: Springer, pp. 238-250.

[40] N. Zampieri, I. Illina, and D. Fohr, "Multiword expression features for automatic hate speech detection," in Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings, 2021: Springer, pp. 156-164.

[41] A. G. d'Sa, I. Illina, and D. Fohr, "Bert and fasttext embeddings for automatic detection of toxic speech," in 2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA), 2020: IEEE, pp. 1-5.
[42] M. A. Bashar, R. Nayak, K. Luong, and T. Balasubramaniam, "Progressive domain adaptation for detecting hate speech on social media with small training set and its application to COVID-19 concerned posts," Social Network Analysis and Mining, vol. 11, pp. 1-18, 2021.
[43] R. Nayak and R. Joshi, "Contextual hate speech detection in code mixed text using transformer based approaches," arXiv preprint arXiv:2110.09338, 2021.
[44] M. Abul Bashar and R. Nayak, "QutNocturnal@ HASOC'19: CNN for Hate Speech and Offensive Content Identification in Hindi Language," arXiv e-prints, p. arXiv: 2008.12448, 2020.
[45] A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, "Hate and offensive speech detection in hindi and marathi," arXiv preprint arXiv:2110.12200, 2021.
[46] G. L. De la Peña Sarracén and P. Rosso, "Convolutional Graph Neural Networks for Hate Speech Detection in Data-Poor Settings," in Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings, 2022: Springer, pp. 16-24.
[47] E. Pronoza, P. Panicheva, O. Koltsova, and P. Rosso, "Detecting ethnicity-targeted hate speech in Russian social media texts," Information Processing & Management, vol. 58, no. 6, p. 102674, 2021.
[48] J. Sánchez-Junquera, P. Rosso, M. Montes, and B. Chulvi, "Masking and bert-based models for stereotype identication," Procesamiento del Lenguaje Natural, vol. 67, pp. 83-94, 2021.
[49] S. Frenda, S. Banerjee, P. Rosso, and V. Patti, "Do linguistic features help deep learning? The case of aggressiveness in Mexican Tweets," Computación y Sistemas, vol. 24, no. 2, pp. 633-643, 2020.
[50] P. Singh and P. Bhattacharyya, "CFILT IIT Bombay@ HASOC-Dravidian-CodeMix FIRE 2020: Assisting ensemble of transformers with random transliteration," in FIRE (Working Notes), 2020, pp. 411-416.
[51] P. Singha and P. Bhattacharyyaa, "CFILT IIT Bombay at HASOC 2020: Joint multitask learning of multilingual hate speech and offensive content detection system," 2020.
[52] B. Jayaraman, T. Mirnalinee, K. R. Anandan, A. S. Kumar, and A. Anand, "Offensive text prediction using Machine Learning and Deep Learning approaches," 2021.
[53] R. Sivanaiah, S. Angel, S. M. Rajendram, and T. Mirnalinee, "TechSSN at semeval-2022 task 5: Multimedia automatic misogyny identification using

deep learning models," in Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 571-574.

[54] D. Thenmozhi, N. Pr, S. Arunima, and A. Sengupta, "Ssn_nlp at SemEval 2020 Task 12: Offense Target Identification in Social Media Using Traditional and Deep Machine Learning Approaches," in Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 2155-2160.

[55] D. Thenmozhi, S. Sharavanan, and A. Chandrabose, "SSN_NLP at SemEval-2019 task 6: Offensive language identification in social media using traditional and deep machine learning approaches," in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 739-744.

[56] R. Priyadharshini, B. R. Chakravarthi, S. Thavareesan, D. Chinnappa, D. Thenmozhi, and R. Ponnusamy, "Overview of the DravidianCodeMix 2021 shared task on sentiment detection in Tamil, Malayalam, and Kannada," in Forum for Information Retrieval Evaluation, 2021, pp. 4-6.

[57] B. R. Chakravarthi, "Multilingual hope speech detection in English and Dravidian languages," International Journal of Data Science and Analytics, vol. 14, no. 4, pp. 389-406, 2022.

[58] B. R. Chakravarthi et al., "Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada," in Proceedings of the first workshop on speech and language technologies for Dravidian languages, 2021, pp. 133-145.

[59] P. K. Kumaresan et al., "Findings of shared task on offensive language identification in Tamil and Malayalam," in Forum for Information Retrieval Evaluation, 2021, pp. 16-18.

[60] J. A. M. Murgado, F. M. Plaza-del-Arco, J. Collado-Montañez, L. A. Ureña-López, and M. T. Martín-Valdivia, "ALIADA: Artificial Intelligence-based language applications for the detection of aggressiveness in social networks," 2022.

[61] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Urena-López, and M. T. Martín-Valdivia, "Comparing pre-trained language models for Spanish hate speech detection," Expert Systems with Applications, vol. 166, p. 114120, 2021.

[62] E. Aldana-Bobadilla, A. Molina-Villegas, Y. Montelongo-Padilla, I. Lopez-Arevalo, and O. S. Sordia, "A language model for misogyny detection in latin american spanish driven by multisource feature extraction and transformers," Applied Sciences, vol. 11, no. 21, p. 10467, 2021.

[63] I. Zupic and T. Čater, "Bibliometric methods in management and organization," Organizational research methods, vol. 18, no. 3, pp. 429-472, 2015.

[64] S. Khanra, A. Dhir, A. N. Islam, and M. Mäntymäki, "Big data analytics in healthcare: a systematic literature review," Enterprise Information Systems, vol. 14, no. 7, pp. 878-912, 2020.

[65] N. Donthu, S. Kumar, and D. Pattnaik, "Forty-five years of Journal of Business Research: A bibliometric analysis," Journal of business research, vol. 109, pp. 1-14, 2020.
[66] H. K. Baker, S. Kumar, and N. Pandey, "A bibliometric analysis of managerial finance: a retrospective," Managerial Finance, 2020.
[67] R. T. Mutanga, N. Naicker, and O. O. Olugbara, "Hate speech detection in twitter using transformer methods," International Journal of Advanced Computer Science and Applications, vol. 11, no. 9, 2020.
[68] N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: a review," IEEE Access, vol. 9, pp. 88364-88376, 2021.
[69] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," PloS one, vol. 14, no. 8, p. e0221152, 2019.