Journal of Scientometric Res. 2022; 11(1):100-111
http://www.jscires.org

Research Article

A Systematic Bibliometric Analysis of Hate Speech Detection on Social Media Sites

Akshaya Gangurde[1], Purva Mankar[1], Deptii Chaudhari[2], Ambika Pawar[1,*]

[1]Symbiosis Institute of Technology, Symbiosis International University (Deemed University), Pune, Maharashtra, INDIA.
[2]Hope Foundation's International Institute of Information Technology, Pune, Maharashtra, INDIA.

ABSTRACT
With the increasing availability of internet facilities for everyone across the globe, the internet plays an integral part in modern communication. People have the ease of contacting others and sharing thoughts and ideas quickly. This has raised an enormous amount of spread of Hate Speech on Online Social Media Sites. This paper aims to provide systematic bibliometric analysis and mappings of existing literature for Hate Speech Detection and to identify the existence of Hate speech-related research. Bibliometric Analysis of Machine Learning and Deep Learning articles in Hate, hostile, and abusive speech is considered. This is accomplished using the SCOPUS database, with tools like VOSViewer, Biblioshiny, and ScienceScape. Explored parameters consist of the document type, most active countries, top journals, relevant affiliations, trending topics, etc. It is observed that the current literature on hate speech is concentrated on a specific philosophy. An unexpected need to rectify this situation was evident from this bibliometric analysis due to recent occurrences of hate speech in the digital world.

Keywords: Hate Speech, Social media, Natural Language Processing, Scopus, Machine Learning, Hostile Speech.

Correspondence
Ambika Pawar
Symbiosis Institute of Technology, Symbiosis International University (Deemed University), Pune, Maharashtra, INDIA.
Email id: ambikap@sitpune.edu.in
ORCID ID: 0000-0003-0842-5751

Received: 07-09-2021;
Revised: 07-06-2022;
Accepted: 19-03-2022.
DOI: 10.5530/jscires.11.1.10

INTRODUCTION

Digital landscapes have globally altered the face of communication. Increased use of this digitized culture, Internet and digital media have become a room for online hate speech or as a digital hate trigger, online hate declaration and cyberbullying.[1] Often this hatred is targeted over to a community or a person of colour, people belonging to different ethnicity, races or to a religious section of people. Hate Speech is considered a blanket for various offensive, abusive, or insulting user-created content. Uncontrolled sharing and posting of content containing Hate Speech is observed on digital platforms which unfortunately, could result in negative psychological effects for certain individuals. Through exploitation of social networking sites as a venue for public interaction is a two-edged sword. Issues about the frequency of hate speech on the internet have recently grown louder. Off-line brutality and volatility may be seen in cyberspace. Organizations, nongovernmental organizations, and broadcasters are advocating for more conversations, as well as more watchdogs and enforcers to combat hateful speech. Strategies to tackle this are brought through by legislation where several social network platforms were required to sign the Hate Speech code. This required various firms to remove Hate content in less than twenty-four hours but even after this only 0.3 of hostile offenders were charged.[2] With an aim to settle the problem, firms seldom rely on the community itself to report the content present. Absence of systematic automatic approaches and data collection on its occasion made the overall process become a complex one. This is where the researchers and scholars initiate their research in Hate Speech identification. One of the significant hurdles in this task is identification of Hate Speech or Abusive Language in 'Hindi' as a natural language and presence of code-mixing (Hindi-English) on online platforms. Code mixed language, recognising false positives, false negatives along with the trends overtime has become a challenge for the research community for detection of Hate Speech.[3]

Natural Language Processing Techniques and development in the Machine Learning Models have brought better insight to this area of research. In a study, researchers were able to identify potential predator activity when it comes to cybergrooming and identification of social media accounts that are responsible for promotion of Hate Speech.[4] Escalated research interest and exploration resulted in regulation of Hate Speech in recent times. Furthermore, an increased demand for research in natural languages other than English is noticed.[5,6] A notable amount of research papers on Hate Speech have been published; a small number of systematic review papers were found during this study. Hence, to the best of our

understanding we endeavoured to give a comprehensive quantitative and qualitative appraisal of the scientific landscape of the publications from 2013 until the present early in the year 2021. The all-inclusive bibliometric analysis also provides standard indicators evaluating the outcomes of publications and analysis of keywords, co-authors, and also citations. Use of visualization techniques allowed better understanding and description about the research work. The study will allow researchers in the area of hate speech to record the noteworthy authors, publications, sources, most relevant keywords, impact of their research work, emerging areas and collaboration opportunities for future research approaches. Statistical evaluation of articles, research areas and publications in this bibliometric analysis would provide a thorough insight for the scientific community. Section 'Related Work' reviews the broad categories in which we classified the previous literature work. Section 'Need for bibliometric analysis' talks about the importance of analysis of data for future proceedings. Section 'Preliminary data collection' presents how the data was procured for statistical analysis. Section 'Bibliometric analysis of first search string' and 'Bibliometric analysis of second search string' deals with analysing the documents retrieved by queries. Section 'Observations and Discussions' gives the learnings drawn from the analysis. Section 'Limitations of Current Work' illustrates the drawbacks of the study. Section 'Conclusion' concludes the analysis and Section 'Future directions 'summarizes the future work.

LITERATURE REVIEW

Extensive previous literature has investigated the growth of Hostile speech towards different communities and groups of victims belonging to various caste, religions, sexual orientations, etc. Over the last ten years, researchers are overseeing an exponential curve on minimizing the presence of online Hate. This serves as a source of inspiration for our work and assists us in associating our study to the current research accessible. Much work has been done on small subject areas of abusive and hostile speech targeted towards minority communities and women. Former work has intensively examined the programmed detection of offensive Internet discourse under an assortment of names, for instance: abusive language, profanity, threats, and socially unacceptable discourse. Many legal issues have revolved around this area of work because of different perceptions of "Hate Speech" as a legal term. The literature associated with Hate Speech can be divided into four categories below.

Legal Assessment

The party-political discussion about the suitable answer to the ever-increasing amount of hate speech on social media has led to a consequent increase in the desire to standardize and even more to automatically identify undesired postings. The legitimate notion of incitement to hatred answers this question by avoiding discrimination against and isolating a target group, thereby guaranteeing the members' acceptance as equivalent in a society - likewise a precondition for democracy. 'Hate speech' is not a legal phrase; the actual appropriate law to this occurrence is, by distinction, particular to each jurisdiction and well-defined. Thus, if the relevant social media post can disrupt public peace since it targets a group, it targets a group established in some country. The need to operationalize this task as an NLP task is crucial.[7] A decision tree is suitable for data annotation along with directions for amateur annotation. Finally, an analysis of how the labels transferred from the decision tree and their annotation can be operationalized as NLP assignments.[8] The subtasks of target group recognition and targeting act recognition can be considered necessary while being annotatable with adequate dependability by non-legally trained individuals. Their findings suggest that it is possible to technically implement this legal task of Hate Speech detection as an automated classification task.[9,10]

Hate Speech against Women

Patriarchal behaviour and other social practices have been transported to the internet, manifesting as misogynistic and sexist remarks, postings, and tweets. This online hate speech against women has a severe outcome in real life. Several legal actions have been lately filed against social networks that fail to prevent the propagation of hate comments targeting individuals.[11] A ground-breaking investigation into online hate speech directed towards women focusing on the distinctions and parallels between misogyny and sexism has started to surface with the up-and-coming technology and research interest directed towards this cause. Discrimination against women seems to be an indicator of a negative attitude towards women.

Experiments have shown that general sexist attitudes conceal a hateful sentiment and, in particular, a misogynistic mindset. Even though sexist humour is usually thought to be guilt-free, numerous researches show otherwise. For example, it emphasizes that sexist jokes are perceived as misogynistic assertions are Frenda S, *et al*.[13]

Furthermore, sexist jokes may contribute to the normalization of sexism or misogyny while also harming the target. Hate speech identification was made using a variety of supervised techniques based on word embeddings. Researchers compared the differences between racist and sexist datasets. They discovered that sexist tweets are more participatory and attitudinal than racist tweets.[14] In this challenging environment, an NLP-based technique can detect the two aspects of patriarchal behaviour, misogyny, and sexism.

Analysing the data and finding good results demonstrate that sexist and misogynistic sentiments are expressions of the patriarchal mentality.

Hate against Multilingual Communities

Social media networks have evolved into a forum where users are free to express their thoughts and feelings, perhaps leading to an increase in hate or abusive communications that must be moderated. Most of the research work is present in English as the prime language.[15] Detection of speech profanity in other languages is still a growing research work. Looking at the diaspora worldwide, researchers have gained interest in exploring Hate Speech in various languages.

From a multilingual standpoint, a supervised technique for hate speech identification is more focused on. Several models have been developed, ranging from feature engineering to neural techniques.

Hate speech encourages prejudice against specific groups and hinders equality, an ongoing problem in every civil society.

Immigrants and women are two groups that are disproportionately targeted.[2,13] Several governments and policymakers are currently attempting to address the issue of immigrants, which has been exacerbated by the refugee crisis and political changes that have occurred in recent years, making the development of tools for the detection and monitoring of such Hate particularly interesting. Furthermore, the work employs a bilingual approach, with data for two widely spoken languages, English, and Spanish, available for training and testing participant systems. The diversity of hate targets and languages creates a unique comparative context regarding the amount of data collected and annotated using the same scheme and the outcomes obtained by participants training their systems on those data. Such a comparative situation may help reveal fresh light on linguistic and communicative behaviour concerning these aims, allowing Hate to be more easily integrated. Speech recognition software for a variety of applications. Experiments in a variety of languages have yielded very encouraging results.

Downgrading Racial Bias

Detection Hate speech, coupled with repressive and abusive language on social media platforms, is part of the current effort, which employs complex algorithms to identify racist or violent speech faster and more accurately without the assistance of humans. On the other hand, machine learning models are prone to inferring human-like biases from the training data used by these algorithms.

There is a strong link between annotators' assessments of toxicity and signals of African American English in contemporary hate speech datasets. Existing automatic detection models overlook an essential factor: context.[16]

Hate speech classifiers are particularly sensitive to group identities such as "transgender," "black," and "gay," which are merely indicators of hate speech in some cases. Because of this bias in annotated training data and the tendency of machine learning models to exacerbate it, AAE text is frequently mislabelled as abusive/offensive/hate speech by existing hate speech classifiers, with a high false-positive rate.[17] Even when there is annotation bias in the underlying training data, a confrontational strategy is to limit the potential of racial bias in hate speech classifiers. When creating a classifier to predict a target attribute, use adversarial training to devalue a protected attribute (AAE dialect) (toxicity).

Necessity of bibliometric analysis

The bibliometric study helps cover the majority of scientific results. It helps analyse published or evaluated articles and citation analysis to look at how those articles influenced subsequent research by others. As a result, this bibliometric evaluation will provide quantitative insights to upcoming researchers in the field of Hate Speech Detection. Hate Speech Detection is also now more effectively possible using Machine Learning and Deeping Learning models. Bibliometric analysis is a great way to get the current trends, understand what has been accomplished in Hate Speech Detection on online platforms, and analyse other literature to optimize the delivery process. The reason of drastic increase in online Hate Speech is the widespread usage of social media which is a powerful instrument for disseminating Hate and abusive language across all digital channels and platforms. Hate Speech is becoming a topic of research in various languages, needing focus to successfully analyse, detect, and neutralize the hostile impacts of propaganda.[18]

METHODOLOGY AND DATA

The Scopus publication database served as our data source for this study. Scopus is a peer-reviewed database of research publications in science, engineering, the arts, social sciences, medicine, technology, and the humanities. Based on detailed bibliometric analysis on the two datasets obtained, the composition of information and the progress of research on the subject of Hate Speech on social media is examined.

The preliminary data collection component of this study is organized as follows: the first section discusses the preliminary data collection procedure and the search technique utilized for data extraction.

The results of bibliometric analysis and data visualization approaches are presented in the following sections. The report

finishes with findings, limits, and recommendations for future research.

Preliminary Data Collection

One of the leading databases for abstracts and citations is the Scopus database by Elsevier. Since 2004, Scopus has been the abode to well-scripted, trustworthy, peer reviews and state-of-the-art research papers that achieve a great level of citations. Scopus has approximately 36,378 documents from roughly 11,677 publishers, of which 34,345 are peer-reviewed journals in top-ranking subject disciplines. It is also developing as a platform that brings researchers, research concepts, and associations together. The data resource for our research work is the standard and reliable Scopus Database.

Our search procedure was broadly split into three sections: data compilation, data mining, data evaluation, and visualization. The time duration for the search was decided to be from 2015 to 2021. For this study, visualization tools used were Scopus and Bibliometrix, an R package utilized to understand the information obtained from Scopus.

Creating the keyword search queries

The main objective of our bibliometric analysis is to map out patterns and trends in the field of Hate Speech detection literature done so far. A preliminary search was engaged using keywords prominently in NLP, Hostile Speech, and Machine Learning paradigms.[1] Keywords are very crucial for searching appropriate research subjects from existing literature. Specific and precise keywords provide a clear-cut illustration of the occurrence of the topic in the same way as our research problem. For our research work, "Hate Speech," "Machine Learning," "Deep Learning," "Social Media," etc., were used. As shown in Table 1, a total of 6 search queries and their outcomes are summarised. In Table 1, search query number 3 resulted in 268 related papers from searching the Scopus Database. Fifty-three related documents were extracted by search string number 1, pinpointed to the research area's language domain, "Hindi." Clear from the results in Table 1, not much study has been performed for Hate Speech Detection in Hindi as a natural language because of various reasons like lack of datasets to train and test the machine learning model and scholars' interest in Hindi dialect.[19,20]

Table 1: Total search queries executed on Scopus.

SN.	Search Query	Results
1.	"Hate Speech" AND "Hindi"	53
2.	("Hate Speech" OR "Hostile Speech") AND ("NLP" OR "Machine Learning)	239
3.	("Hate Speech" OR "Hostile Speech" OR "Abusive") AND ("NLP" OR "Machine Learning" OR "Deep Learning" OR "Neural" OR "LSTM") AND "Social Media"	268
4.	("Hate Speech" OR "Hostile Speech" OR "Abusive" OR "Harm" OR "Toxic") AND ("NLP" OR "Machine Learning" OR "Deep Learning" OR "Neural" OR "LSTM") AND "Social Media"	360
5.	("Hate Speech" OR "Hostile Speech" OR "Abusive" OR "Toxic") AND ("NLP" OR "Machine Learning" OR "Deep Learning" OR "Neuro-Linguistic" OR "LSTM" OR "BERT") AND "Social Media"	282
6.	"Hate Speech" AND ("Twitter" OR "YouTube" OR "Reddit") AND ("Machine Learning" OR "Deep Learning")	129

Preliminary search outcomes using search queries

Via proposed keyword search, data collection was achieved using query strings, which helped retrieve 268 papers for the first query search and 53 documents for the second query search. The primary findings are summarised in Table 1. Table 2 shows the selected queries and the document results from the Scopus database. Table 3 represents the top ten most

Table 2: Selected Search Query.

SN.	Search Query	Result
1.	("Hate Speech" OR "Hostile Speech" OR "Abusive") AND ("NLP" OR "Machine Learning" OR "Deep Learning" OR "Neural" OR "LSTM") AND "Social Media"	268
2.	"Hate Speech" AND "Hindi"	53

Table 3: Top 10 most relevant affiliations.

Affiliations	Articles
Cardiff University	11
Aristotle University of Thessaloniki	7
King Saud University	7
The University of Jordan	7
Dublin Institute of Technology	6
Évry	6
Georgetown University	6
Kempelen Institute of Intelligent Technologies	6
Taibah University	6
University of Central Florida	6

Table 4: Top 10 Cited Documents with year.

Document	Year	Local Citations
Burnap P, 2015, Policy Internet	2015	37
Zhang Z, 2018, Lect Notes Comput Sci	2018	32
Del Vigna F, 2017, Ceur Workshop Proc	2017	25
Kwok I, 2013, Proc Aaai Conf Artif Intell, Aaai	2013	23
Burnap P, 2016, Epj Data Sci	2016	20
Mandl T, 2019, Acm Int Conf Proc Ser	2019	16
Modha S, 2019, Ceur Workshop Proc	2019	16
Macavaney S, 2019, Plos One	2019	14
Pitsilis Gk, 2018, Appl Intell	2018	13
Mandl T, 2020, Ceur Workshop Proc	2020	11

relevant affiliations; Cardiff University is at the top with 11 articles. Table 4 lists the top 10 most cited documents by year. In 2015, 'BURNAP P, 2015, POLICY INTERNET' had a total of 37 citations followed by 'ZHANG Z, 2018, LECT NOTES COMPUT SCI' with a total of 32 citations in 2018.

We are observing the most frequently used keywords in a particular year, Figure 1. shows 'Hate Speech' and 'Machine Learning' having the highest count from 2013 to 2021. Figure 1. Exhibits the top journals publishing hate speech research. Vishwakarma DK is observed to have the most cited documents, with 40 citations in all. In a year-wise document study, the number of cited papers over the years from 2013 to 2021 showed an increase in the curve with a peak of 107 cited documents in the year 2022. Figure 2. Could be referred for the same. Figure 3 illustrates the subject area and the number of documents per subject. Related papers were available primarily from the subject matter of 'Computer Science' with 244 documents followed by 'Engineering' with

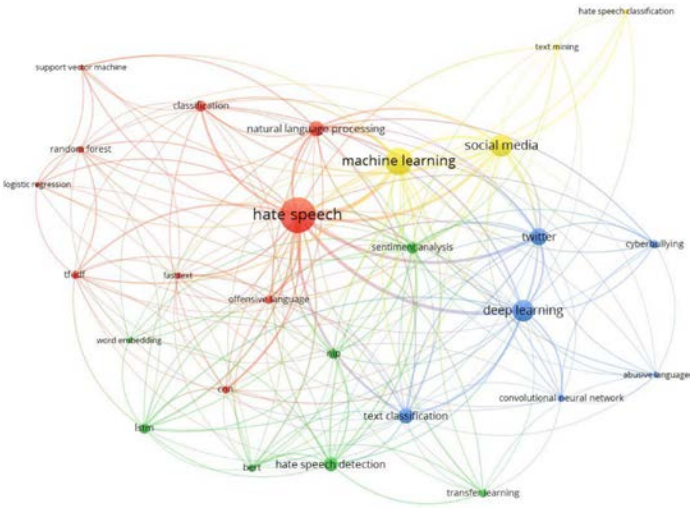

Figure 1: Co-occurrence in Author keywords (Accessed 30 May 2021).

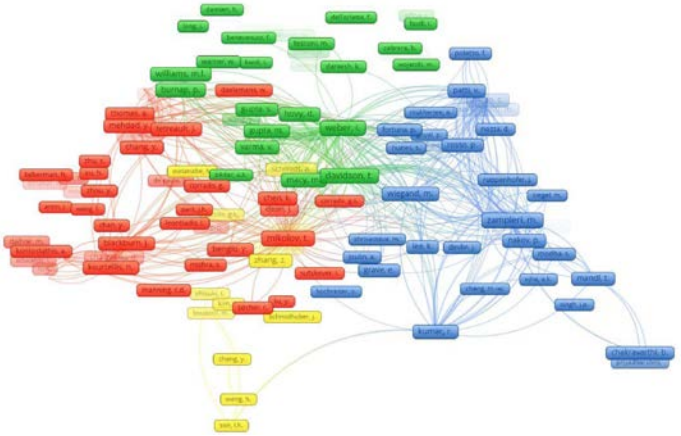

Figure 2: Co-citations of Authors (Accessed 30 May 2021).

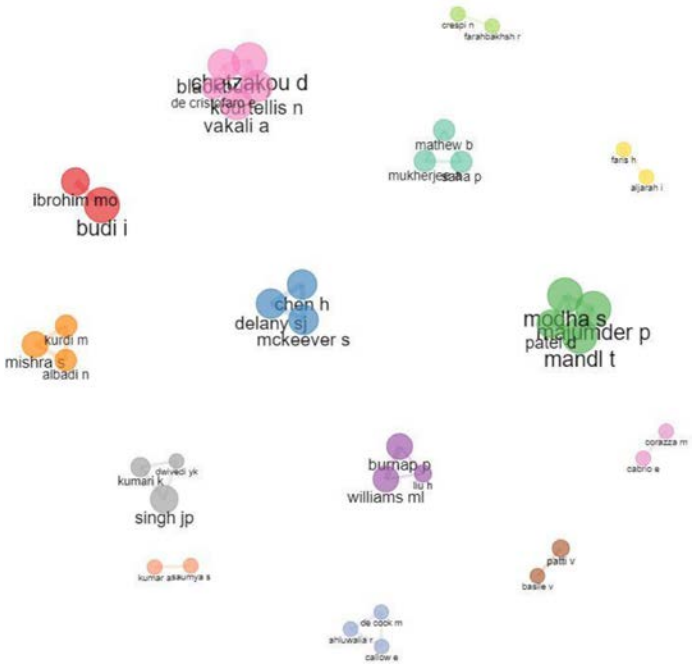

Figure 3: Collaboration Network of Authors (Accessed 30 May 2021).

69 documents. Data visualization tools are used to make data easier to interpret and read trends, patterns, and outliers.[21,22]

Selection of search strings

For our bibliometric analysis, we chose two search strings.

The first query is as follows: ("Hate Speech" OR "Hostile Speech" OR "Abusive") AND ("NLP" OR "Machine Learning" OR "Deep Learning" OR "Neural" OR "LSTM") AND "Social Media."

The first search string has primary keywords like "Hate Speech", "Machine Learning", and "Social Media" and secondary keywords like "Hostile Speech", "Abusive", "NLP", "Deep Learning", "Neural" and "LSTM". This search string was chosen for further bibliometric analysis to gain insights into solving and removing Hate Speech existing on social media sites using different techniques like machine learning.

The second query is as follows: "Hate Speech" AND "Hindi."

The second search string has primary keywords of "Hate Speech" and "Hindi." This search string will provide research publications falling into Hindi as a natural language for hate speech detection. Current research reveals English is one of the pre-distinguished dialects analysed from the point of view of cyberbullying and online hate speech.[14,22] However, there are insufficient widely existing and available datasets in different languages that could pace the growth of research in this field. Other communities might benefit from the removal of Hate from their native dialect. Recent research challenges

include producing massive trustworthy datasets in different languages (Hindi) because online Hate is a prevalent dilemma.

Some overlapping results may be present for the queries executed on the Scopus dashboard.

Bibliometric Analysis of first search string

In this methodical bibliometric learning, discussions were conducted to identify the year-wise trends, quantitatively analyse, define scope, and provide a possibility for better collaboration and exchange of ideas among the research community. Analysis of keywords, collaboration, recognition of various sources of publications, research interest over the years, and co-citation of works was analysed.

The bibliometric analysis is further sectioned into

1. Clustering and Co-occurrences
2. Author, Keyword, and Journal Analysis
3. Statistical Analysis
4. Citation, Document, and Location Analysis

Clustering and Co-occurrences

Figure 1 shows clusters and associations of co-occurrence between author keywords collected using the Scopus Database. Out of 512 keywords, 27 met the threshold of a minimum of five occurrences of a keyword in a document. Total 4 clusters were formed depicted by 4 different colours, with prominent keywords being "Hate Speech," Machine Learning," "Deep Learning," and "Sentiment Analysis" the connection of curved lines shows the researchers' interest in these topics concerning these keywords. Node size determines the author keywords' tally.

Co-citation means the papers, or the authors are cited collectively. For the author co-citation analysis in Figure 2, 154 authors met the threshold criteria, out of 7792 authors, keeping the minimum number of author citations as 20.

Figure 3 determines the collaboration network of Authors. A total of 39 nodes were observed, and disconnected clusters in the Figure are due to a lack of collaboration between the authors. Figure 4 shows the citations according to the source. As observed, 152 documents fit the threshold, out of 152 sources, with the minimum number of documents of a source being 1. The most prominent related item consists of 68 items(21 clusters).

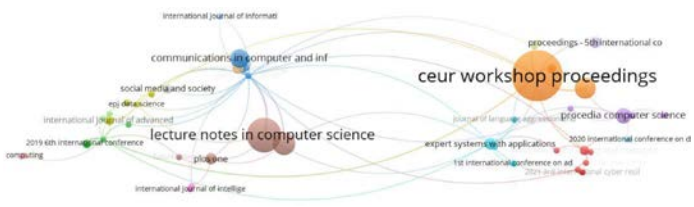

Figure 4: Citations by Source (Accessed 30 May 2021).

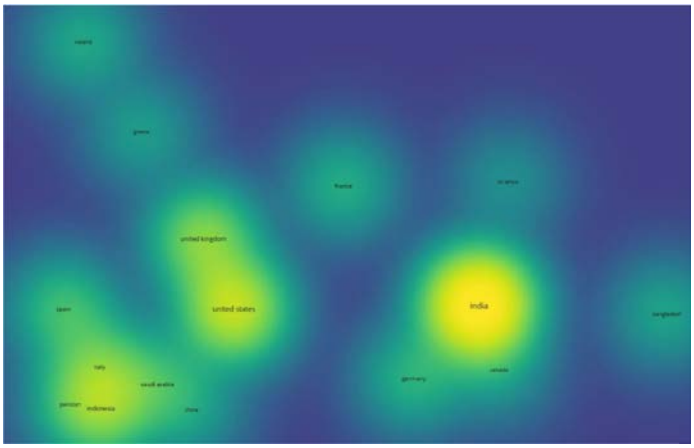

Figure 5: Density Visualization citations by countries (Accessed 30 May 2021).

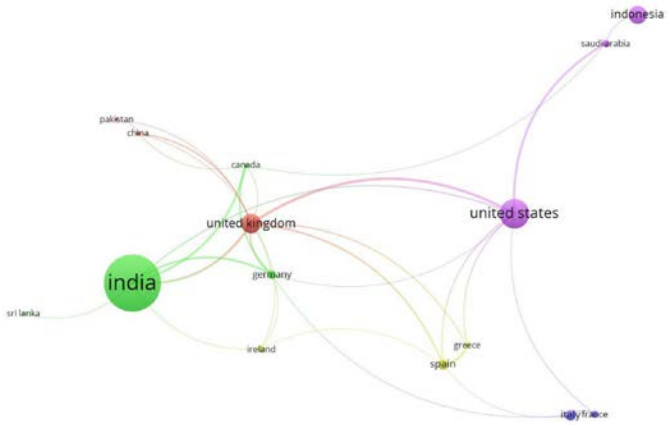

Figure 6: Clusters of co-authorship and countries (Accessed 30 May 2021).

Figure 5 depicts the density visualization of citations by countries. With a minimum 5 number of documents of a country, 16 countries met the threshold out of 54 countries in total. The significant contribution of authors to a journal with countries has been shown in clusters in Figure 6, minimum 5 documents of a country, out of 54 countries, 16 meet the threshold. 15 countries are connected, and 1 is disconnected, which is not shown in the Figure.

Figure 7 shows a coupling map of clusters where the unit of analysis is by documents and the coupling is measured by references. The measure of impact is based on the local citation score. Each cluster is labelled by keyword. 250 units are taken into consideration with a minimum cluster frequency of 5.

Author, Keyword and Journal Analysis

Figure 8 shows authors, keywords, and journals in a network. Most authors have keywords associated with 'Hate Speech,'

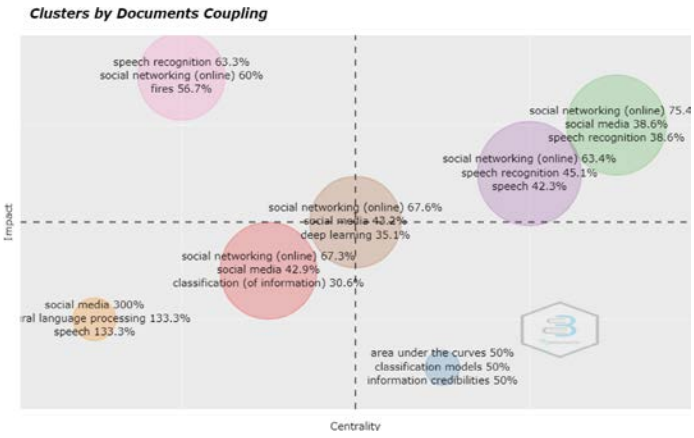

Figure 7: Coupling Map of clusters by Documents (Accessed 30 May 2021).

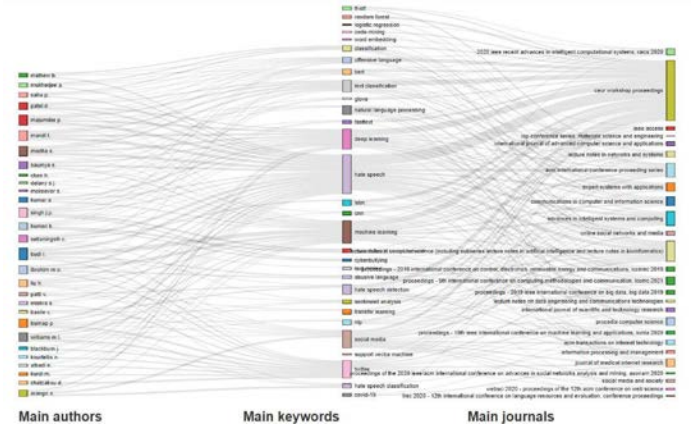

Figure 8: Network of Authors, Keywords and Journals (Accessed 30 May 2021).

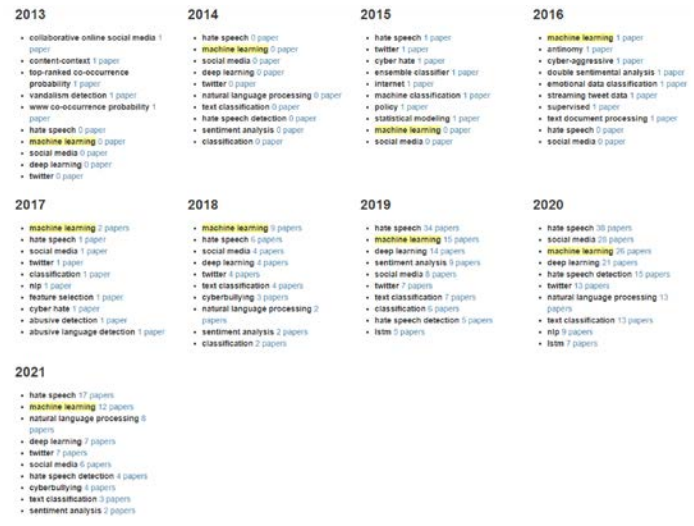

Figure 9: Top Keywords in Year (Accessed 30 May 2021).

'Machine Learning,' 'Deep Learning.' Most keywords are linked to 'CEUR Workshop Proceedings' and 'Lecture notes in Computer Science' (Artificial Intelligence and Bioinformatics).

Figure 10: Top Journals Publishing Hate Speech Research (Accessed 30 May 2021).

Figure 9 shows the top keywords in a particular year, from 2013 to 2020. It is observed that "Hate Speech" and "Machine Learning" keyword count has increased in recent years. Figure 10 shows that the Hate Speech detection research is presented

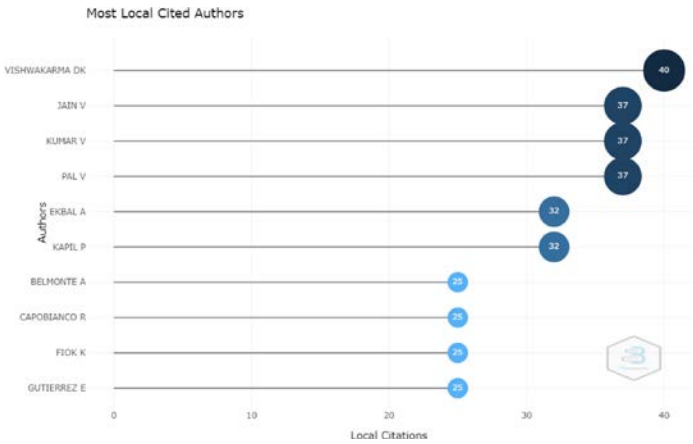

Figure 11: Most Cited Authors (Accessed 30 May 2021).

regularly in the conferences. Few journals like IEEE Access, IEEE Internet Computing, and the International Journal of Scientific and Technology Research have recently published hate speech detection research.

Observing the citations of authors, Vishwakarma DK is recognized to be having the most significant number of locally cited documents, with an overall number of 40 citations in Figure 11.

Statistical Analysis

Figure 12 shows publications about Hate Speech research, which have increased in recent years. This can be attributed mainly to the increased activity of the active social media users on online social media sites and the anonymity provided with it. The development of Machine Learning can also be one of the factors for an increase in Hate and Hostile Speech Detection on Social Media Sites. An increasing number of documents describe the importance of research.

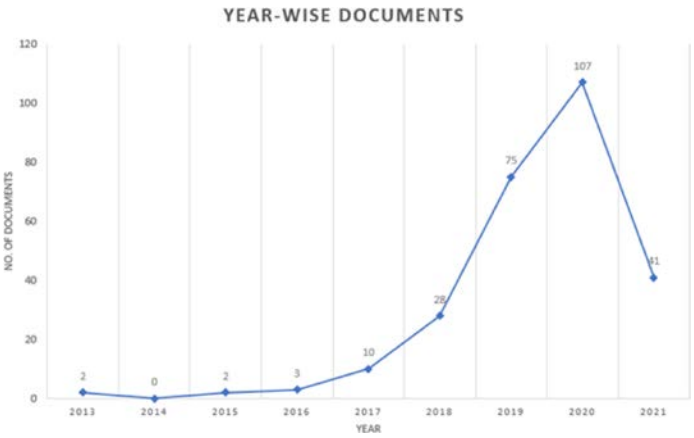

Figure 12: Year-wise Document (Accessed 30 May 2021).

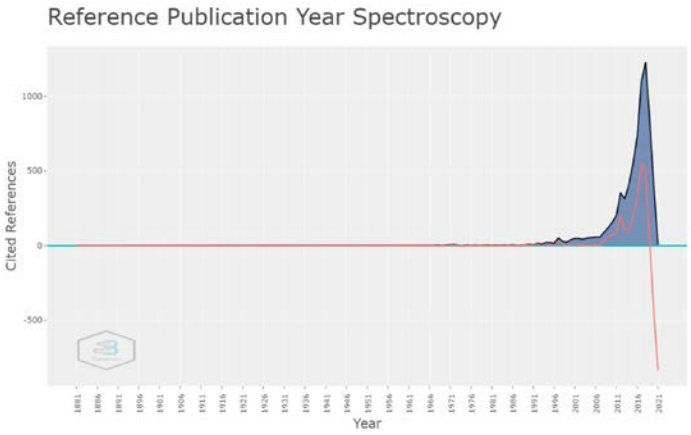

Figure 13: Year-wise cited references (Accessed 30 May 2021).

The cited references per year depict the significance of research in this field. Figure 13 shows a quick rise in citation from the year 2015. This commemorates curiosity held by researchers in the analysis of Hate Speech. As said earlier, it can be attributed to the spread of online Hostile and Hate Speech and the development of machine learning. In 2018, a peak was observed with a statistical number of 1129 references cited in publications.

Figure 14 shows a rapid increase in this research area as many papers have been published from the graph below. We noticed that the number of documents and conference articles in this research section had grown exponentially in these current times. With the increase in user activity and the rise of diverse social media platforms, hate speech has become an additional concern. The quantity of papers is escalating exponentially over the last 4 years. In below Figure 14, in 2020, the maximum number of articles were published.

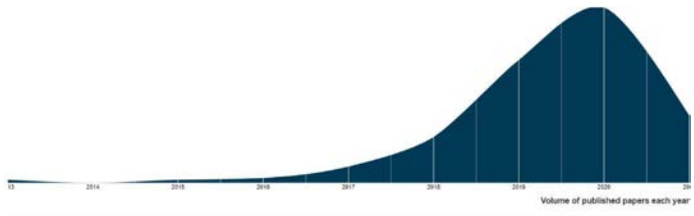

Figure 14: The volume of papers overtime (Accessed 30 May 2021).

In the literature published from 2013 to 2021, 19 subject areas were included, referred to, and worked upon hate speech. Computer Science has the highest number of documents (244). Since most hate speech is prevalent in the digital world, Computer Scientists and researchers are working on various tools to recognize and eliminate such content. Engineering has 69 documents of such research work. As shown in Figure 15, it demonstrates the distribution of subject areas in the research area.

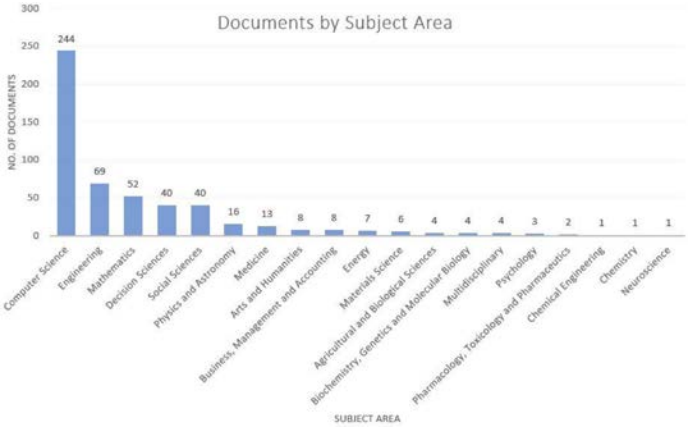

Figure 15: Documents by subject area (Accessed 30 May 2021).

Citation, Document, and Location Analysis

Figure 16,17 depicts that conference papers are a popular choice of the researcher to publish Hate Speech Detection research. Twenty-five percent of the extracted documents are published as journal articles.

Word dynamics determines the word growth of cumulative occurrences of top 10 field keywords. "Social Networking

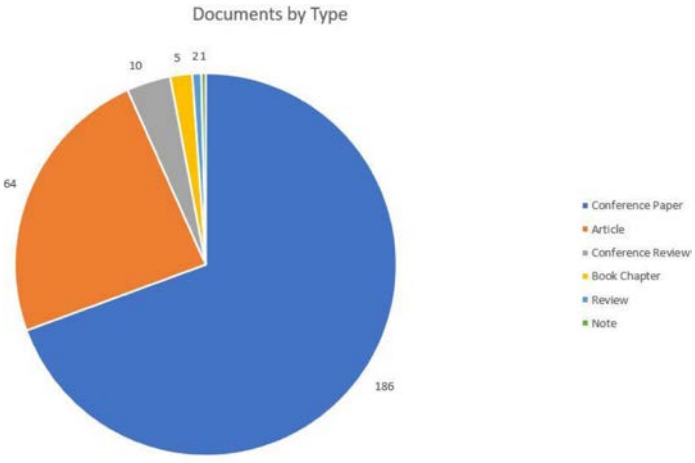

Figure 16: Documents Type (Accessed 30 May 2021)

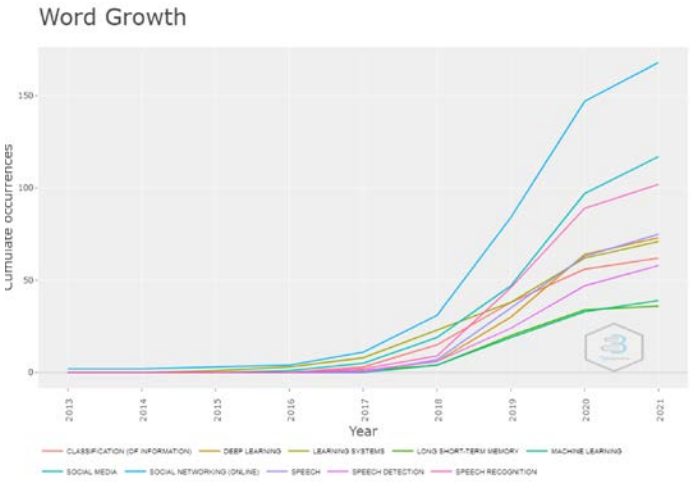

Figure 17: Word Dynamics (Accessed 30 May 2021).

(Online)" has a total of 168 occurrences in the year 2021 and "Social Media" has a total of 117 occurrences in the year 2021, and "Speech recognition" having a cumulative occurrence of 102 in the year 2021 as well. Topics like these have the most considerable term frequency, indicating a trending fashion in this area of research.

Figure 18 shows the connection between various countries, with the USA, India, and UK showing the most collaboration. India and the USA lead in Hate Speech Detection research,

Figure 18: World Collaboration Map (Accessed 30 May 2021).Bibliometric Analysis of second search string

having many studies revolving around Hostile Speech detection, NLP, and Machine Learning topics.

In this systematic bibliometric analysis of the second string, discussions were conducted to discover the year-wise trends and define scope and subject areas that further probe into depth. Evaluation of the most commonly used keywords, global collaboration of authors, and recognition of various sources of publications, research interest over time, and co-citation were carried out in order to learn more about this research gap, motivating many more researchers to enter this field and contribute to it.

The bibliometric analysis is further sectioned into

1. Statistical Analysis
2. Author, Keyword, and Journal Analysis

Statistical Analysis

Hate Speech Detection has not been explored more in the Hindi language. There has been an insufficient number of articles and documents published in this particular area of Hate

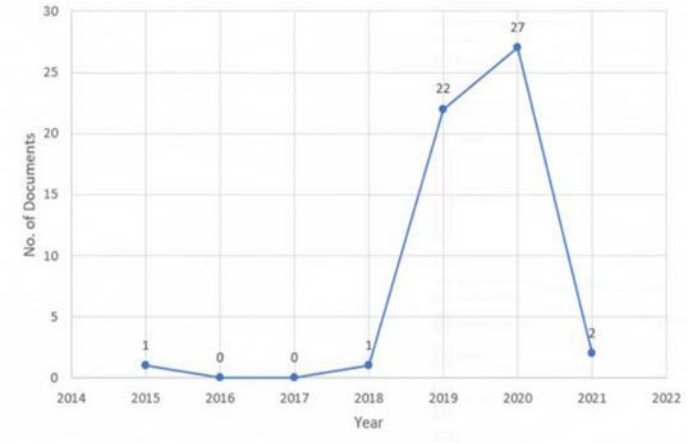

Figure 19: Documents by Year (Accessed 31 May 2021)

Speech research. As evident from Figure 19, there has been a recent pick of pace with detection language being "Hindi." The year 2020 has seen the highest number of publications of 27. Working on this research gap would bring our fruitful contribution to the table.

Figure 20 shows the gradual increase from 2018 to 2020 in this field of work, with "Hindi" being the natural language of detection for Hate Speech. With the rise in the "Machine Learning" and "Deep Learning" techniques, it has been easy

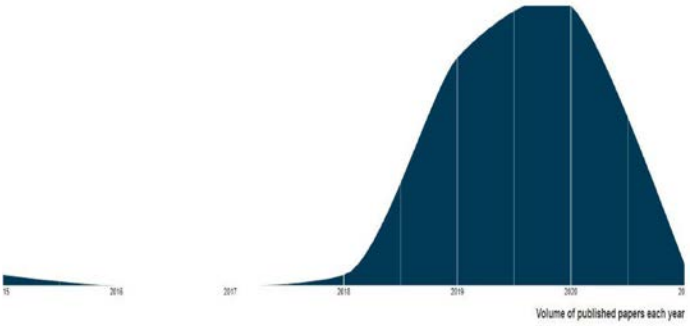

Figure 20: The volume of papers overtime (Accessed 31 May 2021).

and helpful to detect and remove Hate Speech from Online Social Media platforms. More significant studies are taking place with the development of different NLP techniques.

Author, Keyword and Journal Analysis

Figure 21 shows the contribution of authors to several journals with the most used keywords like "Hate Speech," Deep Learning," "Text Classification," etc. The most significant

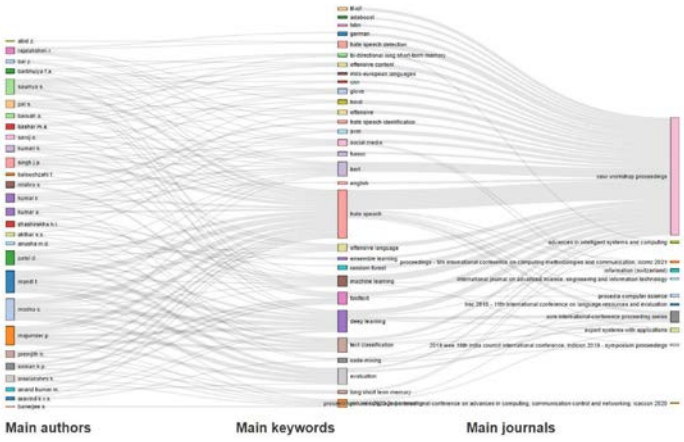

Figure 21: Network of Authors, Keywords and Journals (Accessed 31 May 2021).

number of authors publishing documents to "CEUR Workshop Proceedings."

n Figure 22, the period from 2015 to 2021, the contribution of researchers in the field of Machine Learning and Hate Speech Detection has shown an increase. The researchers became

Figure 22: Main Keywords overtime (Accessed 31 May 2021).

interested in Hindi as a natural language as the area of research and code-switching. Different techniques for detecting Hate in the Hindi language were used like BERT, text classification, fastText, deep learning, etc.

Figure 23 is a thematic map showing development degree (density) versus relevance degree (centrality) for publications titles. Only 10 multilingual publications were observed

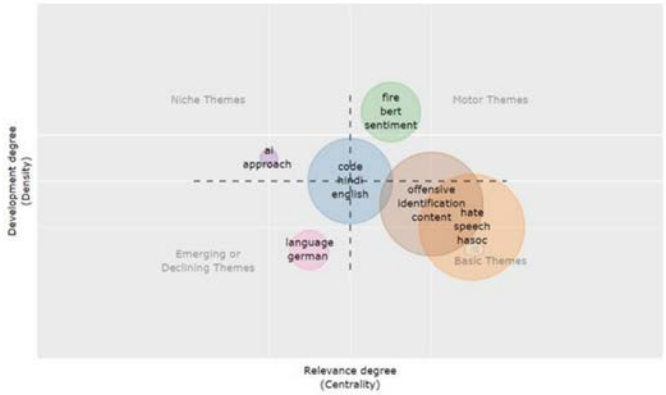

Figure 23: Thematic Map (Accessed 31 May 2021).

overlapping offensive identification content and hate speech. Identification of 10 published papers in pure Hindi Language and 8 code mixed Hindi-English papers is observed. Showing significantly less research has been done for multilingual languages.

Figure 24 shows all 53 documents retrieved, illustrates the subject area-wise classification for this research area. Maximum research has been carried out in computer science,

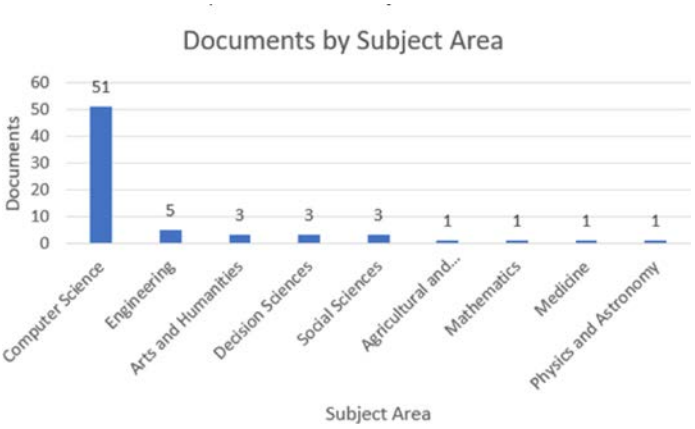

Figure 24: The document by Subject Area (Accessed 31 May 2021).

51 documents belong to "Computer Science" subject area followed by "Engineering" and "Arts and Humanities" which is clearly indicated from the figure.

Observations and Discussions

From 2013 to 2021, the goal of the bibliometric analysis was to examine the trends and progress of publications linked to Hate Speech Detection on social media. Hate and Hostile Speech detection is a multidisciplinary area of research. Language modelling and semantics, social behavioural analysis, and psychological analysis are some of the subject areas pursued by researchers with Machine Learning, Deep Learning and

Natural Language Processing for Hate Speech detection. Machine Learning practices dominate the current. To enhance the research value Natural language processing is applied with Machine Learning. Deeply analysing and prospecting the graphs and their trends, we can confirm positive movements. Figure 12 and Figure 19 indicate increased year- wise growth in publications in this field. The graph for word dynamics depicted in Figure 17 illustrates the use of machine learning, speech detection, deep learning, and long short-term memory in a smaller number of studies, indicating a possible research gap which can be explored further by fellow researchers to investigate different possible ways to detect and dissolve online hate with technologies like natural language processing and machine learning. Conferences are the favoured publication approach by the researchers working towards this domain. India and the USA lead in Hate Speech research which is evident from the concentrated areas in the world map presented in Figure 18. From the second search string, we can draw a conclusion as not much study is present in Hindi as a dialect. But there has a growing trend observed in statistical analysis in Figure 19, with 2020 having the highest of only 27 publications. There was a steep rise seen between the years 2018 and 2020, which is a positive sign for growing research in this gap.

According to the findings of this study, knowledge on inciting online hatred is rising at an accelerating rate in the computer engineering, decision sciences and medicine sector, and it is currently one of the most active and fastest-growing previous research and research activity in the social sciences as well as the technology domain. The quantity of articles in 2020 demonstrates that there is a great deal of interest in the issue among social critics and scholars. The researchers of previous articles are certain that the 2021 January Capitol Hill Storming event, as well as the subsequent social media prohibition and termination of such profiles, will spark more curiosity and study in this field. It should be mentioned that both developed countries and countries trying to develop are actively conducting research on this issue. India, a country rich in variety and home to a plethora of different religions, is increasingly concerned about strategies to combat hate speech.

Hate speech analysis literature may be seen focusing on numerous elements such as hate speech directed at diverse groups of individuals in society, such as cultural minorities, women, and people of various sexual orientations. At this moment, the need for more advanced tools and strategies to solve this issue is critical.

Scientists have concentrated on identifying posts that express a disparaging view about a recognised personality of the country in more published findings. Intentionally posting anything with hostile intent against an individual has far-reaching social, political, and economical consequences. As a result, it's critical to categorise every social media post that could lead to a negative perception of a person. The studies showed a collaborative work between Computer Vision Techniques such as Optical Character Recognition (OCR) which extracts texts from images and Natural Language Processing (NLP) to focus on features of digital platforms such as comments of the related post, context of the post and particularly memes which are images of people with text. They typically encompass a wide range of content and a graphic that shows entertaining things to the audience.

Understanding Memes present a new set of hurdles for researchers, as they need concurrent visual and textual comprehension. The post content (primary text) and connected comments, in addition to Memes, offer significant details about the post's goal. The user's formal response, known as the post content or main content, is used in conjunction with Memes to communicate his or her standpoint. Likewise, comments are the text sentences written through other people in response to a person's content.

In recent years, machine learning, deep learning, and natural language processing (NLP) techniques have attracted academics' attention to work on detection difficulties in order to achieve a more efficient result with fewer downsides. Hate Speech detection in Hindi dialect lies in a research gap due to the lack of datasets available. From Figure 23, this thematic map reveals that approaches like "AI Approach" and "Sentiment analysis" lie in the niche theme for providing the solution to the problem of online hate. Researchers are starting to also work on code mixed datasets with the increasing use of code-mixed style on online social media platforms.

Limitations of current work

There are different research databases like Web of Science, PubMed, etc. But for this Bibliometric study, only the SCOPUS database is considered. It is not comprehensive and does not include all scholarly articles. Another drawback is that there is a chance that publications will overlap. Research papers were extracted based on time criteria chosen between 2013 and 2021. Despite these restrictions, the patterns and trends observed in this study are unlikely to be affected.

CONCLUSION

This study used bibliometric analysis to look at hate speech on social media. For research publications in the Scopus database, we conducted an analysis from 2013 to 2021. Based on our findings, it appears that research in the field of hate speech identification is growing. Our data show that cooperation between the United States, India, and the United Kingdom are common in this sector of research. Vishwakarma DK, Jain V, and Kumar V are among the most often mentioned

authors in the area. The authors with affiliations to Cardiff University, Aristotle University of Thessaloniki, and King Saud University contributed the majority of the publications. The keyword analysis refers to social media studies of hate speech aimed at exposing issues such as prejudice against women, racial bigotry, and hatred directed at multilingual populations.

The goal of this work was to use bibliometric analysis to show the current state of Hate Speech detection research on social media. Our findings revealed that such experiments are becoming more refined and are being featured in top publications, as well as cross disciplinary journals involving linguistics and politics, education among other disciplines. In a research of social media channels, Twitter came out on top as the most popular digital site and one of most normally employed language in social networking sites exploration was English.

As social media has become an increasingly important part of our lives, study into the analysis and detection of profanity appears to be more important than ever. The novelty of this work is that it effectively demonstrates a thorough bibliometric analysis of the topic. Bibliometric tools such as VOSviewer and Biblioshiny are used to create a mind map, co-occurrence, co-citations, Sankey plot, and world cooperation map.

Future Directions

This paper aims to provide a clear understanding of the characteristics and research potential for detecting hate speech in social media. This analysis summarizes prior work in this field of research for academics interested in contributing to this field of research. To assist other studies and research projects in a strong direction by using all available analysis and data to refer to future trends. Based on a thorough examination, we believe that the research on Hate Speech Detection on social media will be fruitful in computer science. We anticipate that additional detailed studies on hate speech identification on online social media platforms in Hindi will be conducted.

Comparisons of different topic models in identifying dominating themes and issues in a particular research subject might be noteworthy and valuable in future work.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Mladenovic M, Ošmjanski V, Stankovic SV. Cyber-aggression, Cyberbullying, and Cyber-grooming: A Survey and Research Challenges. ACM Computing Surveys. 2022;54(1):1-42. doi: 10.1145/3424246.
2. Vashistha N, Zubiaga A. Online multilingual hate speech detection: experimenting with Hindi and English social media. Information. 2021;12(1):5. doi: 10.3390/info12010005.
3. Banerjee S, Chakravarthi BR, McCrae JP. Comparison of pretrained embeddings to identify hate speech in Indian code-mixed text. In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). IEEE Publications; 2020;21-5.
4. Kim YJ, Qian L, Aslam MS. Bibliometric analysis on workplace cyberbullying: study protocol. F1000Research. 2021;10(225):225. doi: 10.12688/f1000research.51495.1.
5. Mandl T, Modha S, Kumar MA, Chakravarthi BR. Overview of the hasoc track at fire 2020: hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In Forum for Information Retrieval Evaluation. 2020;29-32.
6. Que Q, Sun R, Xie S. Simon@HASOC 2020: detecting hate speech and offensive content in German language with BERT and ensembles. InFIRE [working notes] 2020;283-9.
7. Zufall F, Zhang H, Kloppenborg K, Zesch T. Operationalizing the legal concept of'Incitement to Hatred'as an NLP task. arXiv preprint arXiv:2004.03422. 2020.
8. Chaudhari DD, Pawar AV. Propaganda analysis in social media: a bibliometric review. Information Discovery and Delivery. 2021;49(1):57-70. doi: 10.1108/IDD-06-2020-0065.
9. Mishra Dr R. Are we doing enough? A bibliometric analysis of hate speech research in the selected database of Scopus.
10. Twitter.com. Updating our rules against hateful conduct. Twitter safety. Available from: https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html [cited 29/3/2022].
11. Silva L, Mondal M, Correa D, Benevenuto F, Weber I. Analyzing the targets of hate in online social media. In: Tenth international AAAI Conference on WEB and Social Media. 2016.
12. Sap M, Card D, Gabriel S, Choi Y, Smith NA. The risk of racial bias in hate speech detection. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019:1668-78.
13. Frenda S, Ghanem B, Montes-y-Gómez M, Rosso P. Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. Journal of Intelligent and Fuzzy Systems. 2019;36(5):4743-52. doi: 10.3233/JIFS-179023.
14. Facebook.com. Facebook Community Standards: Hate speech. Available from: https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/
15. Singh P, CFILT BP. IIT Bombay at HASOC 2020: joint multitask learning of multilingual hate speech and offensive content detection system. In FIRE [working notes]. 2020;325-30.
16. Tontodimamma A, Nissi E, Sarra A, Fontanella L. Thirty years of research into hate speech: topics of interest and their evolution. Scientometrics. 2021;126(1):157-79. doi: 10.1007/s11192-020-03737-6.
17. Rodriguez A, Argueta C, Chen YL. Automatic detection of hate speech on Facebook using sentiment and emotion analysis. In 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE Publications. 2019;169-74.
18. Sreeram G, Sinha R. A Novel Approach for Effective Recognition of the Code-Switched Data on Monolingual Language Model. InInterspeech. 2018:(1953-7).
19. Gupta V, Sehra V, Vardhan YR. Hindi. English code-mixed hate speech detection using character level embeddings. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC). IEEE Publications. 2021;1112-8.
20. Sreelakshmi K, Premjith B, Soman KP. Detection of hate speech text in Hindi-English code-mixed data. Procedia Computer Science. 2020;171:737-44. doi: 10.1016/j.procs.2020.04.080.
21. Qazvinian V, Rosengren E, Radev D, Mei Q. Rumor has it: identifying misinformation in microblogs. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011;1589-99.
22. Gaikwad M, Ahirrao S, Phansalkar SP, Kotecha K. A bibliometric analysis of online extremism detection. Library Philosophy and Practice. 2020;1:1-6.