applied sciences

MDPI

Article

A Bibliometric Analysis of Text Mining: Exploring the Use of Natural Language Processing in Social Media Research

Andra Sandu [1], Liviu-Adrian Cotfas [1,*], Aurelia Stănescu [2] and Camelia Delcea [1]

[1] Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, 010552 Bucharest, Romania
[2] Department of Management, Bucharest University of Economic Studies, 010552 Bucharest, Romania
* Correspondence: liviu.cotfas@ase.ro; Tel.: +40-771269599

Abstract: Natural language processing (NLP) plays a pivotal role in modern life by enabling computers to comprehend, analyze, and respond to human language meaningfully, thereby offering exciting new opportunities. As social media platforms experience a surge in global usage, the imperative to capture and better understand the messages disseminated within these networks becomes increasingly crucial. Moreover, the occurrence of adverse events, such as the emergence of a pandemic or conflicts in various parts of the world, heightens social media users' inclinations towards these platforms. In this context, this paper aims to explore the scientific literature dedicated to the utilization of NLP in social media research, with the goal of highlighting trends, keywords, and collaborative networks within the authorship that contribute to the proliferation of papers in this field. To achieve this objective, we extracted and analyzed 1852 papers from the ISI Web of Science database. An initial observation reveals a remarkable annual growth rate of 62.18%, underscoring the heightened interest of the academic community in this domain. This paper includes an n-gram analysis and a review of the most cited papers in the extracted database, offering a comprehensive bibliometric analysis. The insights gained from these efforts provide essential perspectives and contribute to identifying pertinent issues in social media analysis addressed through the application of NLP.

Keywords: text mining; natural language processing; social media; bibliometric; Biblioshiny

check for updates

Citation: Sandu, A.; Cotfas, L.-A.; Stănescu, A.; Delcea, C. A Bibliometric Analysis of Text Mining: Exploring the Use of Natural Language Processing in Social Media Research. *Appl. Sci.* **2024**, *14*, 3144. https://doi.org/10.3390/app14083144

Academic Editors: Marija Brkić Bakarić and Maja Matetic

Received: 15 March 2024
Revised: 5 April 2024
Accepted: 8 April 2024
Published: 9 April 2024

1. Introduction

The preceding decades have seen evident progression in technology, which has generated numerous advantages and considerably simplified people's lives along with setting the path to an ever-changing future.

The economy, communication, health, education, entertainment, and many other developed industries have all experienced radical shifts as a result of this rapid evolution. The emergence of new jobs, the development of automated and far more efficient equipment, the reduction of bureaucratic processes, and the facilitation of information access can be listed as some of the key benefits that have completely impacted the way people relate to one another, communicate, and carry out their daily activities.

Additionally, real-time communication and the ease of worldwide connectivity served as the foundations for the enormous growth of social networks, which are experiencing an unprecedented level of popularity today. There is hardly anyone nowadays who does not have an account on at least one of the leading social media platforms, such as Facebook, Instagram, or Twitter, whereas only a few years ago, people were hesitant to use these networks and barely heard of them. These social media networks have become crucial tools that are accessible to almost everybody via a wide range of devices, therefore making it simpler than ever to share text, video, and multimedia content. They serve as the best place for promoting certain events or goods, disseminating news, sharing opinions, and engaging in debates surrounding particular subjects. The evolution of social media, along with its impact in different areas, such as on organizations and leaders [1], businesses [2], and marketing [3], has been widely explored in the scientific literature.

Machine learning, a branch of artificial intelligence, has the main goal of providing techniques that allow machines to learn without the need for explicit programming, making it possible for computers to learn from data and past experiences and identify trends as well as forecast results with little assistance from individuals. Despite its modest beginnings in the 19th century, with algorithms involving linear regression, decision trees, and K-means, the field progressed throughout the years due to technological advancements and increased access to wide amounts of data, which resulted in the development of more sophisticated and effective techniques such as deep neural networks. Nowadays, voice assistance, personalized suggestions, medical diagnosing, treatment solutions, and many other essential fields use machine learning, a discipline that is constantly and rapidly expanding and developing. In this context, Callier [4] discusses the role of machine learning in evolutionary studies, Dejan et al. [5] provide an assessment of various machine learning models for sensor-extracted data, and Telikani et al. [6] provide a survey on evolutionary machine learning.

Natural language processing focuses on the in-depth understanding of a text, with its main goal being to achieve an appropriate interpretation and to be able to respond and generate human language. Although it started from simple approaches such as morphological analysis, nowadays, with the help of techniques based on machine learning, it manages to learn complex models, perform sentiment analysis, translate text automatically, provide virtual assistance, build chatbots, extract data, provide summaries, and perform numerous other tasks. The progress in this area has been synthesized in various studies from the field, such as those by Ortiz-Garces et al. [7], Chang [8], Hirschberg [9], Zhang et al. [10], Jiang et al. [11], and more.

As a result, natural language processing and machine learning are tightly associated and extremely helpful in numerous different domains. The use of social networks to extract users' messages associated with certain events and opinions, followed by their analysis through natural language processing, brings to light crucial facts. As a result, this field has attracted a large number of researchers who have included these subjects in their publications. The outcomes of these research investigations can be extremely helpful for developing strategies, comprehending current problems in society, increasing the general level of happiness, or even combating the spread of fake news and disinformation. In this context, Pandey et al. [12] underline the use of natural language processing for sentiment analysis in social media marketing, while Al-Saif et al. [13] explore the role of emotions in social media detecting. Furthermore, the work of Boon-Itt and Skunkan [14] discusses the public perception of the COVID-19 pandemic on Twitter by addressing issues related to sentiment analysis and topic modeling.

Thus, this current study aims to bring the area of natural language processing in social media research into focus by choosing a relevant set of articles from this field; presenting the materials and methods used in detail; and then performing a comprehensive analysis of various perspectives, including sources, authors, an examination of the existing literature, and a summary of the first 10 most cited articles, as well as an investigation of words, universities, and countries, by employing a multitude of graphs, indicators, and visual representations. Considering the purpose of our paper, bibliometric analysis was chosen as the investigation technique. As Block and Fisch [15] noted in their research, contrary to a review analysis whose scope is to provide a summary of a particular research field in terms of content, with an accent on the most important findings, bibliometric analysis focuses more on the structure of a particular field, making an emphasis on its development. Furthermore, as we think that it is important to also have a view of NLP in social media research, we decided to enhance the analysis by providing a review of the top 10 most cited papers. The choice for this option is grounded on the idea that the papers with the highest number of citations have represented an important base for other studies in the field, succeeding in reaching a broader audience compared to the other papers included in the dataset. By proceeding in this manner, the advantages of a review analysis have been combined with the advantages of a bibliometric analysis, providing a more comprehensive view of NLP use in social media research.

As noted, the main purpose of this work is to present the insights gained from the analysis and to discover new trends, hidden information, and results that can lay the foundations for future political, economic, medical strategies, etc. This manuscript also aims to answer questions such as those listed below:

- Q1. Which are the most cited articles in the area of natural language processing in social media research, and what insights can be uncovered?
- Q2. Who are the most prolific authors in the studied domain, and what is their country of origin?
- Q3. What valuable information can be brought to the fore, regarding the collaboration of authors in the area of natural language processing in social media research?
- Q4. How can the authors' production over time be characterized?
- Q5. What facets come to the forefront through the performed analysis of words?

In view of the preceding statements, this paper is organized in six sections, as follows: Introduction (Section 1)—an overview of the background information and the topics addressed; Materials and Methods (Section 2)—a presentation of the tools employed and the process of selecting and filtering the dataset; Dataset Analysis (Section 3)—a detailed analysis of the dataset; Discussions (Section 4)—a presentation of the most important insights found following the analysis; Limitations (Section 5)—drawing attention to limitations and boundaries; Conclusions (Section 6)—a presentation of the findings along with suggestions for future work.

2. Materials and Methods

The purpose of this section is to bring the steps followed in our analysis to the forefront, along with the tools and methods used for gathering and exploring the dataset, with the main intention of providing valuable and noteworthy results for interested readers around natural language processing in social media research.

As a starting point of the discussion, each bibliometric analysis involves two distinct steps, highlighted in Figure 1. The first stage is known as dataset extraction, and the second stage is associated with bibliometric analysis, both presented below, in the ensuing sub-sections.

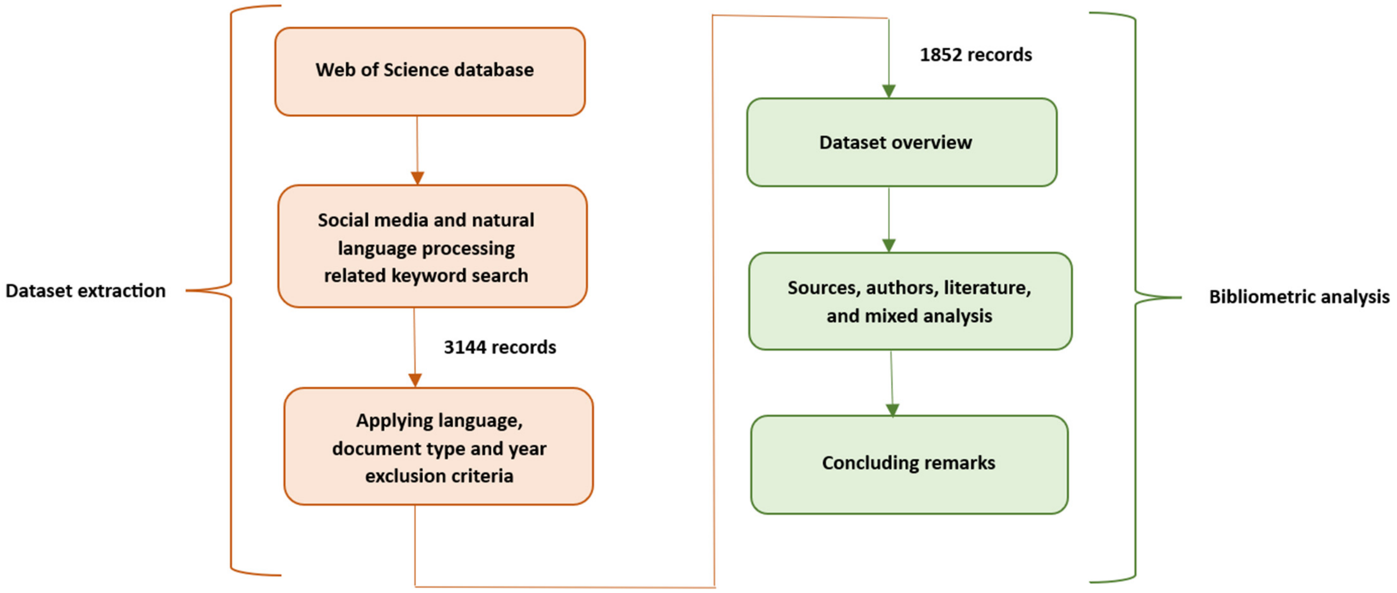
Dataset extraction
Web of Science database
Social media and natural language processing related keyword search
3144 records
Applying language, document type and year exclusion criteria
1852 records
Dataset overview
Sources, authors, literature, and mixed analysis
Concluding remarks
Bibliometric analysis

Figure 1. Bibliometric analysis steps.

2.1. Dataset Extraction

Since the main purpose of bibliometric analysis is to focus on the development of a particular field, the extraction of the dataset is truly important, and the authors should pay high attention to this when performing the process of selection.

The research tool harnessed in the present manuscript for extracting the collection of papers is the Web of Science database [16], also referred to as WoS. This instrument has a pivotal role for bibliometric analysis, facilitating access to a wide variety of academic publications across a multitude of domains.

The decision to exclusively use the WoS database is supported by multiple arguments, including the ones exemplified below:

- It has a preeminent and influential position within the academic community compared to other databases (IEEE, Scopus).
- It grants users personalized access to sources, following a subscription-based approach (in the case of this article, the researchers had full access—all ten indexes). The papers by Liu [17] and Liu [18] addressed the importance of having full access to sources when performing bibliometric analysis.
- It is up to date and includes a broad spectrum of scientific articles from multiple domains and journals, being highly recognized by academics. The articles written by Bakir et al. [19], Cobo et al. [20], and Mulet-Forteza et al. [21] provide more insight into this matter.
- It offers users the possibility to import data extracted from WoS in row format into the popular R tool (4.3.2), Biblioshiny (3.2.1).
- In most of the academic papers employing bibliometric analysis [22–27], published over the years on different subjects, researchers extracted data from the WoS database. This serves as proof for considering WoS as the main database for the data collection step in our analysis.
- This bibliometric analysis includes individual analyses, such as the most cited articles, which is something that would have generated issues when shaping the list of the most cited articles according to the number of citations if we used several databases.

However, we are aware of the limitations and impact of our decision on the final dataset, and we agree that if we had used multiple databases in our analysis, the results might have been slightly different.

All the exploratory steps and filters applied on the WoS database are explained in detail in Table 1, along with the queries and count for each.

As can be observed, the first data selection step involved three distinct queries that were used to search for specific keywords in either the titles, abstracts, or authors' keywords.

The primary goal of the first query was to extract the documents that contained the "social media" keyword, which generated a set of 113,396 papers. This ensured that the extracted corpus was directly related to social media and addressed topics, issues, and challenges specific to this field, being relevant to the goal of the article.

The next query had the purpose of delving further into the subject of "natural language processing", and this term was searched for in titles, abstracts, and keywords. The result revealed that 39,766 papers had been written around this trending subject, demonstrating its influence and relevance within the scientific community.

The last step was to combine the findings garnered in the aforementioned queries, certifying that the extracted set included only pertinent papers that address the social media domain within the context of natural language processing. This merging process generated a unique dataset composed of 3144 documents.

The second exploratory step focused on excluding papers written in languages other than English, since English is considered a universal language within the scientific community. This serves as proof of the widespread understanding of English among researchers, leading to the majority of studies being written exclusively in this language. However, after applying this criterion, the dataset saw a slight decrease, reaching 3112 documents, a small difference compared to the initial amount of 3144 papers, which reinforces the previous hypothesis that English is the most prevalent language among academics.

The third exploratory step involved selecting only the papers marked as "article" from the data collection set in the WoS database. It should be noted that in the type of document noted as "article", WoS includes all new and original works that are considered citable [22]. Therefore, research papers, brief communications, technical notes, chronologies, full papers, and case reports (presented like full papers) that were published in a journal and/or presented at a symposium or conference were included in this category [28]. Consequently, this decision had a major impact on the number of documents, which dropped to almost half, from 3112 to 1866 pieces.

Finally, the fourth exploratory step eliminated the articles written in 2024, since they were still in progress at the moment of conducting the analysis. This ruling was based on the desire to have a well-defined and precise timespan, specifically focusing on the exclusive analysis of articles published in the years 2010–2023. The final corpus processed through bibliometric analysis consisted of 1852 articles.

Table 1. Data selection steps.

Exploration Steps	Filters on the Web of Science	Description	Query	Query Number	Count
1	Title/Abstract/Author's Keywords	Contains specific keywords related to social media in title/abstract/author's keywords	((TI = (social_media)) OR AB = (social_media)) OR AK = (social_media)	#1	113,369
		Contains specific keywords related to natural language processing in title/abstract/author's keywords	((TI = (Natural_Language_Processing)) OR AB = (Natural_Language_Processing)) OR AK = (Natural_Language_Processing)	#2	39,766
		Contains specific keywords related to social media and natural language processing in title/abstract/author's keywords	#2 AND #1	#3	3144
2	Language	Limit to English	(#3) AND LA = (English)	#4	3112
3	Document Type	Limit to Article	(#4) AND DT = (Article)	#5	1866
4	Year Published	Exclude 2024	(#5) NOT PY = (2024)	#6	1852

2.2. Bibliometric Analysis

Once the data selection process was performed, the next step on the list involved performing bibliometric analysis.

With the intention of extracting and populating tables, as well as obtaining graphs along with visual representations useful for this scientific research, the R tool Biblioshiny [27,29] was used. This outstanding instrument is truly valuable and effective in bibliometric analyses, as through the multitude of features and options provided, it brings crucial information about the imported data collection set to the spotlight, exposes insights and trends, and reveals hidden details in the field of natural language processing in social media research, contributing to the development and improvement of strategies, together with establishing an online world with its main objectives of increasing the level of happiness as well as protecting users' privacy and reliability to the greatest extent.

Based on Figure 2, one can notice five different perspectives through which the data were assessed: dataset overview, sources, authors, papers, and mixed analyses.

The first facet shed light on the main elements included in the dataset, the key details about the data, and primary information related to the authors and their collaboration.

Source analysis brought facts about the journals to the forefront, as well as their impact and growth based on the number of publications throughout the selected timespan.

Dedicated to author examination, the third perspective unveiled insights about the most prolific authors and their contribution to the area of natural language processing in social media research, the most relevant affiliations, countries, scientific production, and collaboration map analysis.

The fourth facet focused on paper analysis and presented key details about the most globally cited documents, along with a deep investigation into the words extracted.

The last one, mixed analysis, highlighted crucial insights about the associations among various elements by making use of three-field plot representations.

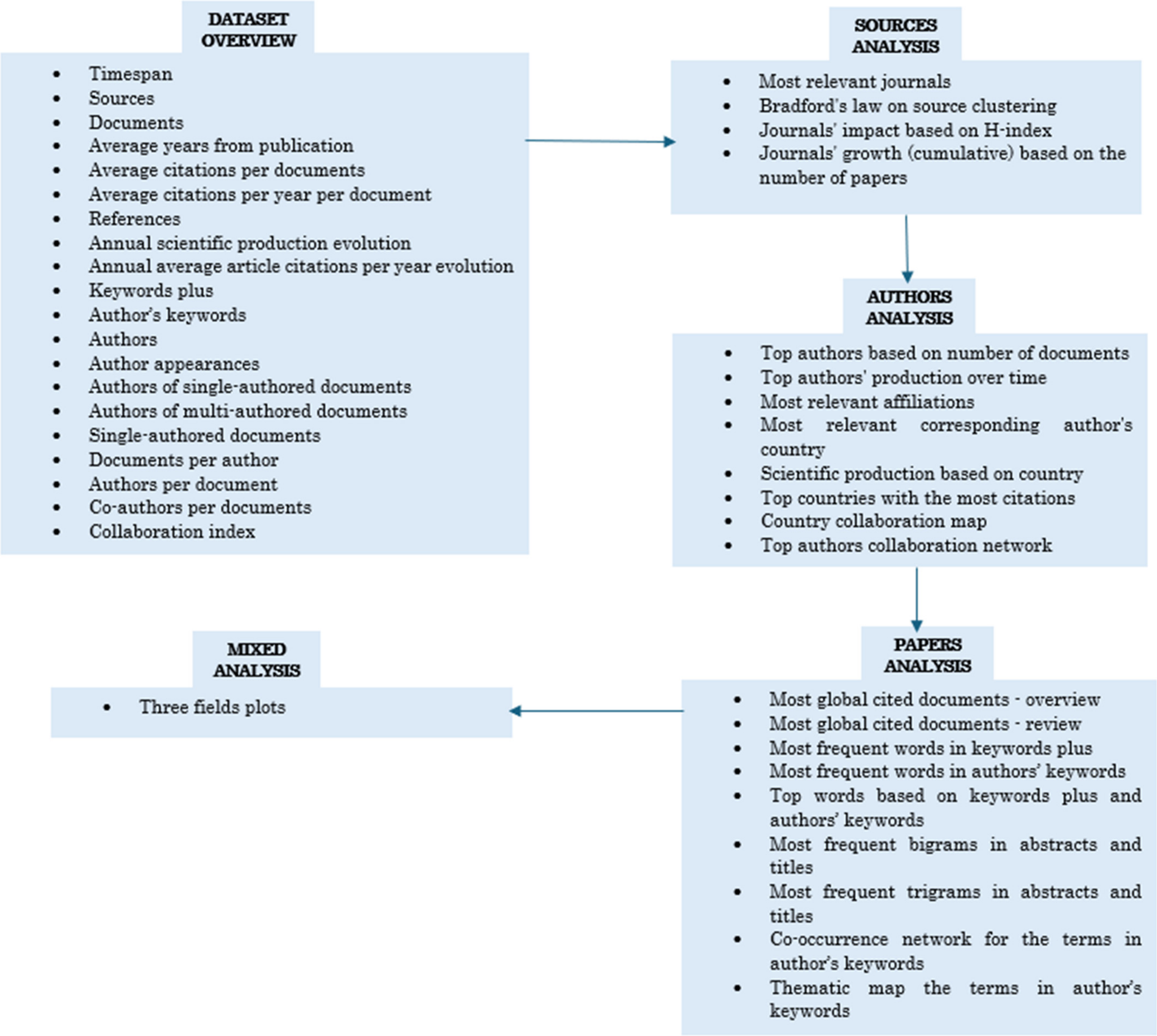
DATASET OVERVIEW
Timespan
Sources
Documents
Average years from publication
Average citations per documents
Average citations per year per document
References
Annual scientific production evolution
Annual average article citations per year evolution
Keywords plus
Author's keywords
Authors
Author appearances
Authors of single-authored documents
Authors of multi-authored documents
Single-authored documents
Documents per author
Authors per document
Co-authors per documents
Collaboration index
SOURCES ANALYSIS
Most relevant journals
Bradford's law on source clustering
Journals' impact based on H-index
Journals' growth (cumulative) based on the number of papers
AUTHORS ANALYSIS
Top authors based on number of documents
Top authors' production over time
Most relevant affiliations
Most relevant corresponding author's country
Scientific production based on country
Top countries with the most citations
Country collaboration map
Top authors collaboration network
PAPERS ANALYSIS
Most global cited documents - overview
Most global cited documents - review
Most frequent words in keywords plus
Most frequent words in authors' keywords
Top words based on keywords plus and authors' keywords
Most frequent bigrams in abstracts and titles
Most frequent trigrams in abstracts and titles
Co-occurrence network for the terms in author's keywords
Thematic map the terms in author's keywords
MIXED ANALYSIS
Three fields plots

Figure 2. Bibliometric analysis facets.

3. Dataset Analysis

The collection of articles related to natural language processing in social media research, extracted and filtered during the previous step, will be meticulously analyzed in this section from several points of view, like sources, authors, citations, words, affiliations, collaborations, and several more.

3.1. Dataset Overview

Table 2 brings key details about the data collection set into the limelight, comprising 1852 articles, disseminated across 697 sources, during a relatively large period of time—specifically, 14 years between 2010 and 2023—highlighting the fact that area of natural language processing in social media research seems to capture the attention of researchers over a considerable timespan and simultaneously proving its significance within the scientific community.

The relevance and current importance of the study performed by researchers in this article is further supported by the low value of 2.98, registered as the average years from publication, suggesting that most of the articles written around the studied domain are recent papers.

Furthermore, the academic significance of this field is also deduced from the recorded values of 12.76 for average citations per document and 2.467 for average citations per year per document.

In terms of references, a substantial number was noted—more precisely, 67,244.

Table 2. Main information about the data.

Indicator	Value
Timespan	2010:2023
Sources	697
Documents	1852
Average years from publication	2.98
Average citations per documents	12.76
Average citations per year per document	2.467
References	67,244

Furthermore, in terms of paper type, the following categories have been included in the "article" type provided by WoS: *article* (1715 papers), *article: book chapter* (19 papers), *article: data paper* (5 papers), *article: early access* (102 papers), and *article: proceedings papers* (11 papers).

The annual scientific production evolution is reflected in Figure 3. One can easily notice an evident rising tendency in published articles that address the topic of natural language processing in social media research, experiencing an annual growth rate of 62.18%.

Of the total amount of 1852 articles selected in the analysis, only one article was published in 2010. In light of its humble beginnings, there is a slight annual rise in the article count until 2018, when it reached a number of 70 documents. Henceforth, this domain registered astonishing growth, an expansion in the production of articles in this field, attaining a significant number of 537 published papers in 2023, highlighting both the relevance of the field and the enthusiasm and burgeoning interest of scientists.

Figure 4 presents a fluctuating pattern of average citations per year, marked by variable numerical values between 0.1 and 8.5, suggesting strong visibility of the articles within the scientific community, covering the years 2010 through 2023. Regarding the evolution of the number of citations per year, it should be mentioned that, as expected, the papers published in the years closely to the end of the year of analysis have a smaller time window in which the papers could be cited, which explains the drop in citations registered for the 2020–2023 period. Therefore, by considering a reduced time window for 2010–2020, in terms of citations, it can be observed that the average citations per document indicator equals 29.57 citations per document—visibly higher than the value for the same indicator provided in Table 2 for the entire dataset (12.76 citations/document). Furthermore, the value of the average citations per year per document is 4.353 for the selected 2010–2020 period, even higher in this case than the value recorded for the same indicator in the case of the entire dataset, namely 2.467 citations/year/document.

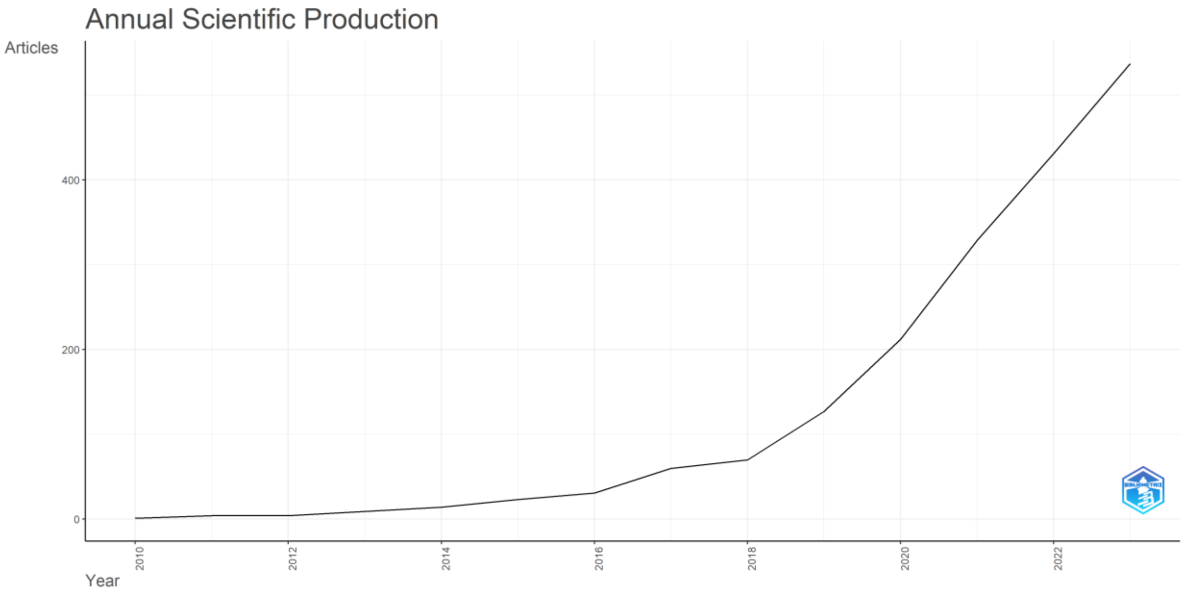

Figure 3. Evolution of annual scientific production.

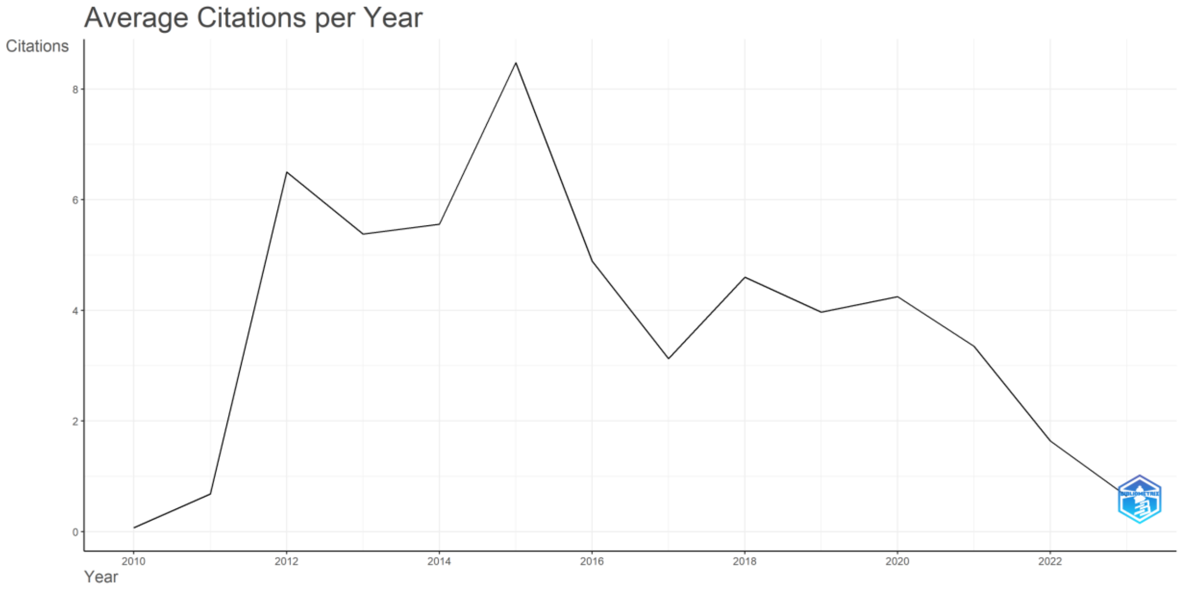

Figure 4. Evolution of annual average article citations per year.

Moreover, as depicted in Figure 4, a spike in data can be observed for 2015 and 2012 in terms of the evolution of average article citations per year. Considering the papers published in the above-mentioned years, it can be observed that the spike in the annual average article citations for 2015 can be attributed to the paper authored by Ravi et al. [30], which registered a total citation value of 675 and a total citation per year value of 67.50. As for 2012, the papers of Reyes et al. [31] and Montoyo et al. [32] resulted in 153 and 140 citations, respectively, significantly contributing to the spike in the evolution of average article citations per year for the mentioned year.

Table 3 brings to the fore details regarding the document contents. In terms of keywords plus, also referred to as index terms, a significant number of 1607 entries were registered, and, through the division of this value by the total number of documents extracted for the analysis, namely 1852, one can obtain an average of 0.87 such terms per document. These numerical values may indicate the utilization of a concentrated vocabulary within the scientific papers in the analyzed domain.

There were also 4311 author keywords reported, with these mentioned terms averaging 2.33 per document.

Table 3. Document contents.

Indicator	Value
Keywords plus	1607
Author's keywords	4311

Data about the authors are revealed in Table 4. A discrepancy that proves dense collaboration within the area of natural language processing in fake news research is noticed by comparing the total number of identified authors, namely 6238, with the total number of articles collected for bibliometric analysis, 1852.

The increased value of 7675 registered for the author appearances indicator suggests the fact that among the 6238 distinct authors spotted, there are also researchers involved in the authorship of multiple scientific works in this field.

To further extend the discourse, although it is obvious that the authors who addressed this subject in their research opted for collaborations with other members—a hypothesis also proven by the increased value of 6178 registered by the single-authored documents indicator—it is noteworthy to state that within the collection of papers selected, 60 single-authored documents are noticed as well.

Table 4. Authors.

Indicator	Value
Authors	6238
Author appearances	7675
Authors of single-authored documents	60
Authors of multi-authored documents	6178

Table 5 delves deeper into the ongoing discussion about author collaboration. Based on the data provided below, and along with the knowledge gathered until this point, it is not a surprise that collaboration in the domain of natural language processing in social media research is increased, a fact also suggested by the low value obtained for the single-authored document index. By drawing a comparison between 60 and 65, namely, the indicators for authors of single-authored documents and the single-authored documents themselves, it can be confirmed that the scientists who decided to publish articles as sole authors in this domain accomplished it for an average of 1.08 documents.

These statements are further supported by the decreased value recorded for the documents per author index, more specifically 0.297, a decreased value obtained since the number of authors is greater than the number of documents (6238 versus 1852). As a result, it generated an average of 3.37 authors per document, 4.14 co-authors per document, and 3.46 for the collaboration index.

Table 5. Author collaboration.

Indicator	Value
Single-authored documents	65
Documents per author	0.297
Authors per document	3.37
Co-authors per document	4.14
Collaboration index	3.46

3.2. Sources

Figure 5 places the spotlight on the top 19 most relevant journals. In the process of obtaining the chart below, an exclusion criterion was considered. Only the journals with at least 16 articles published around the area of natural language processing in social media research were included.

As can be noticed, the foremost position is held by the popular journal *IEEE Access*, which recorded a significant value of 92 published articles in the analyzed area, followed closely by the *Journal of Medical Internet Research* with 85 documents. For the entire list, please consult Figure 5.

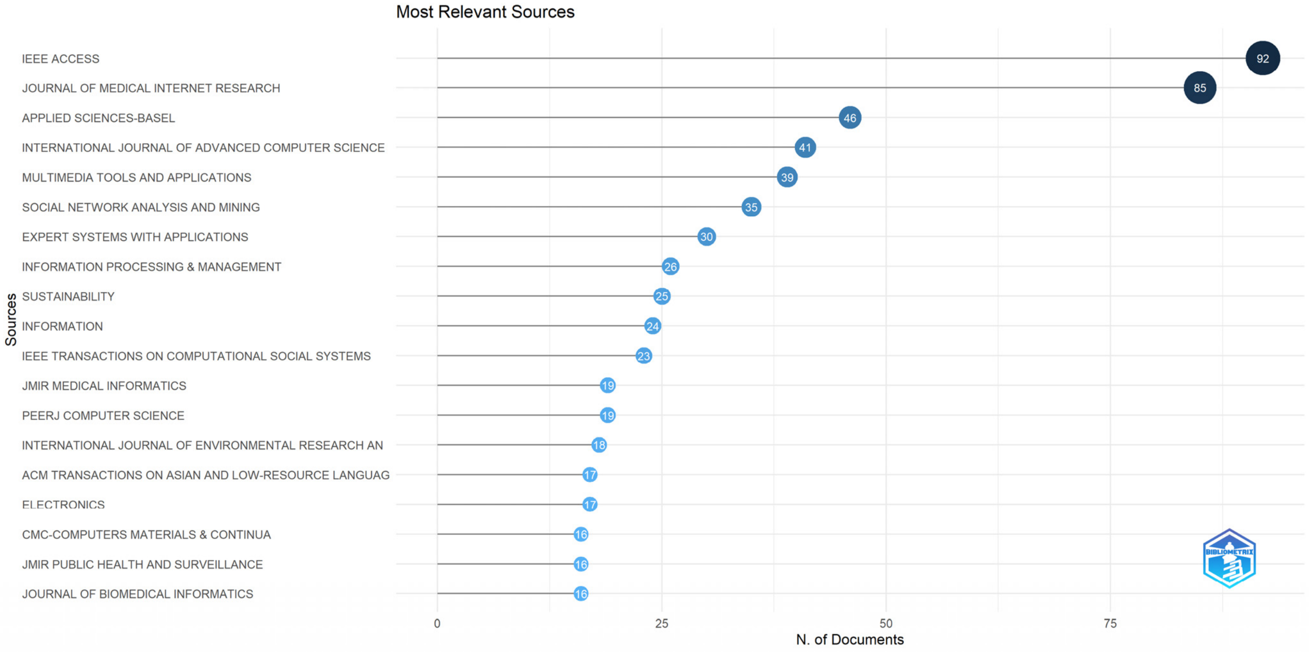

Figure 5. Top 19 most relevant journals.

Figure 6 depicts the core sources by using Bradford's law, a popular theory which divides three distinct zones based on the significance of the sources. According to Bradford's Law, the identified sources are divided into three main categories based on the number of published papers [33,34]. If the proportion of the articles in each category would be one-third of all the articles, then the number of journals in each group would be proportional, with 1:n:n^2 [26,33,34]. The journals placed in the core (Zone 1) are very significant and highly cited, the journals in the center (Zone 2) are moderately significant, while the journals in the external zone (Zone 3) comprise a larger number of less significant journals.

That being said, in the case of our analysis, the most significant and highly cited journals, found in Zone 1, are listed here in order: *IEEE Access; Journal of Medical Internet Research; Applied Sciences-Basel; International Journal of Advanced Computer Science and Applications; Multimedia Tools and Applications; Social Network Analysis and Mining; Expert Systems with Applications; Information Processing & Management; Sustainability; Information; IEEE Transactions on Computational Social Systems; JMIR Medical Informatics; PeerJ Computer Science; International Journal of Environmental Research and Public Health; ACM Transactions on Asian and Low-Resource Language Information Processing; Electronics; CMC-Computers, Materials & Continua; JMIR Public Health and Surveillance; Journal of Biomedical Informatics; Frontiers in Public Health* (please see Figure 6).

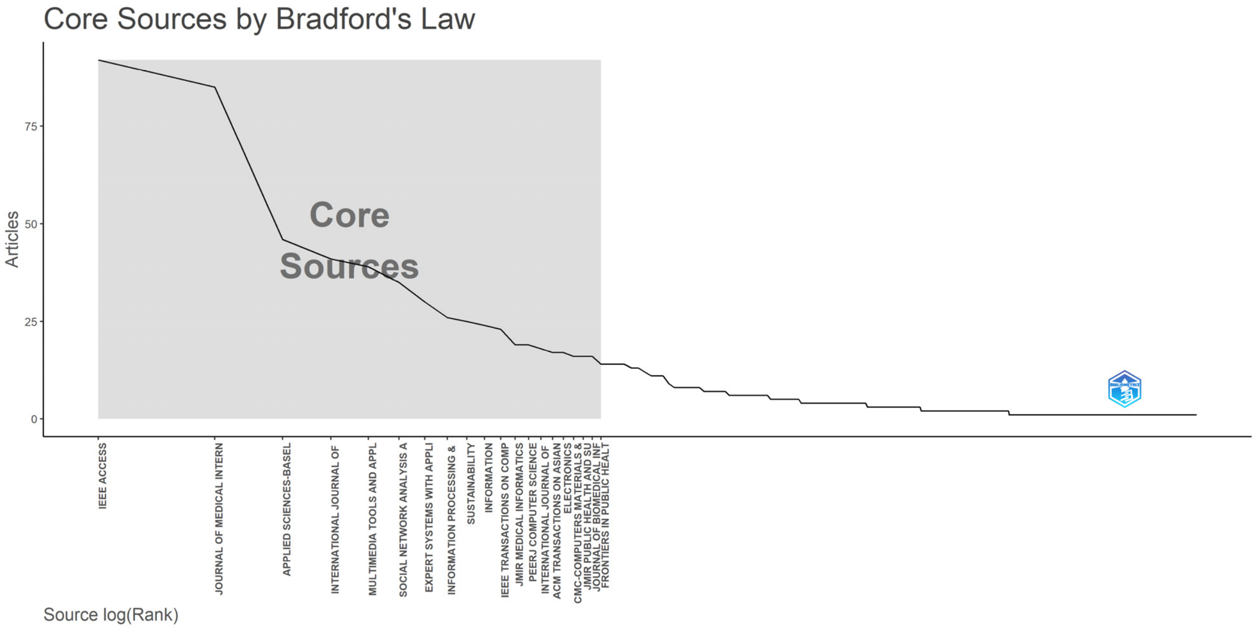

Figure 6. Bradford's law on source clustering.

Figure 7 brings to the foreground the journals' impact based on the h-index, a metric that proves the significance of a source by measuring the total count of published articles in the studied area, which register H-citations. The leading position is held by the *Journal of Medical Internet Research* with 20 citations. The second place is occupied by *IEEE Access*, which gathered 17 citations, followed closely by *Information Processing & Management* with 15 citations. For the entire list, kindly refer to Figure 7.

Furthermore, it is not surprising that the journals with substantial impact based on h-index analysis can also be found in Zone 1, delimited above using Bradford's law.

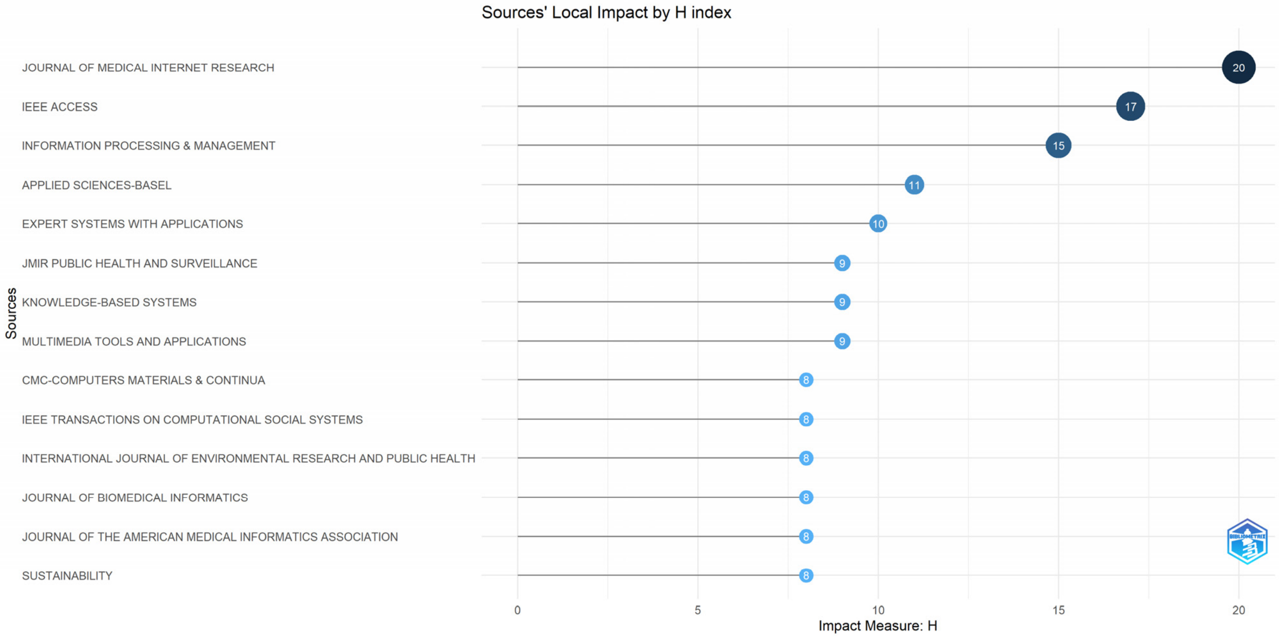

Figure 7. Journals' impact based on the h-index.

Sources' production over time is captured in Figure 8. As anticipated, based on previous findings, the journals that mark the most pronounced growth are *Applied Sciences-*

Basel, *IEEE Access*, *International Journal of Advanced Computer Science and Applications*, *Journal of Medical Internet Research*, and *Multimedia Tools and Applications*.

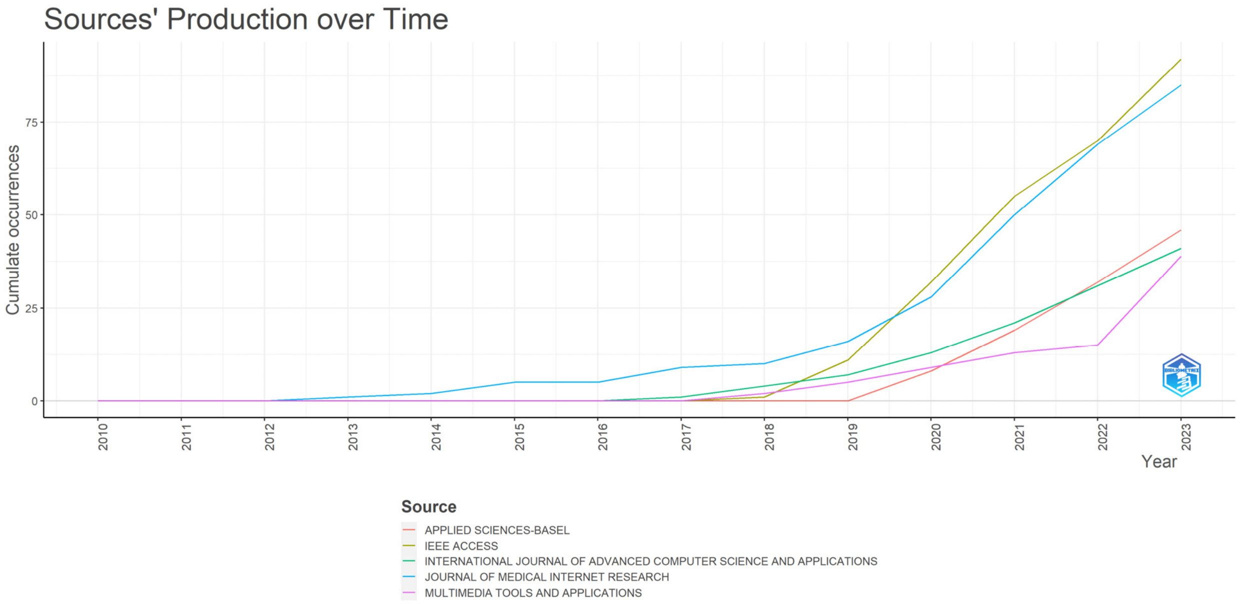

Figure 8. Journals' growth (cumulative) based on the number of papers.

3.3. Authors

The most prolific scientists selected through a criterion that only considers authors with at least six published articles around natural language processing in social media research are brought into focus in Figure 9.

The foremost position is held by Sarker A with 28 published papers, followed, with a considerable difference, by Gonzalez-Hernandez G with 14 articles, and O'Connor K with 13. For the entire list, direct your attention to Figure 9.

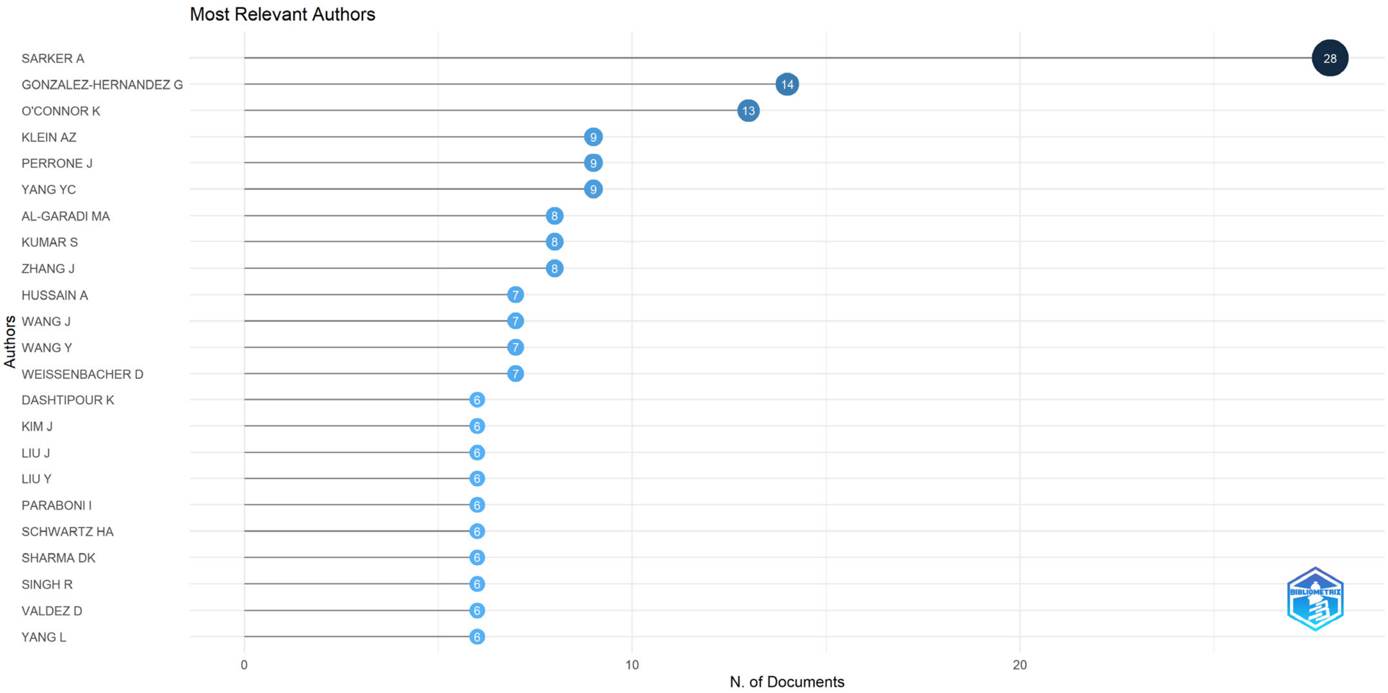

Figure 9. Top 23 authors based on the number of documents.

In terms of author production over time, referring to Figure 10, natural language processing in social media research seems to be a topic that has attracted the attention of researchers for quite a long time.

With a rather modest beginning in 2015, only three articles were published (Sarker A—2 papers and Gonzalez-Hernandez G—1 paper), but from that moment onward, in the succeeding period, a substantial annual increase in the number of publications is observable.

An outstanding increase is registered starting with the years 2020–2021, when scientists directed special attention to this topic, publishing more than one article per year. The increase may be justified by the occurrence of some events such as the COVID-19 pandemic or the outbreak of wars, which triggered a multitude of users' reactions on social media. Natural language processing is a valuable tool for understanding people's feelings, along with their level of happiness in connection with certain happenings, while social media is the best place to find opinions and conflicting debates around a subject.

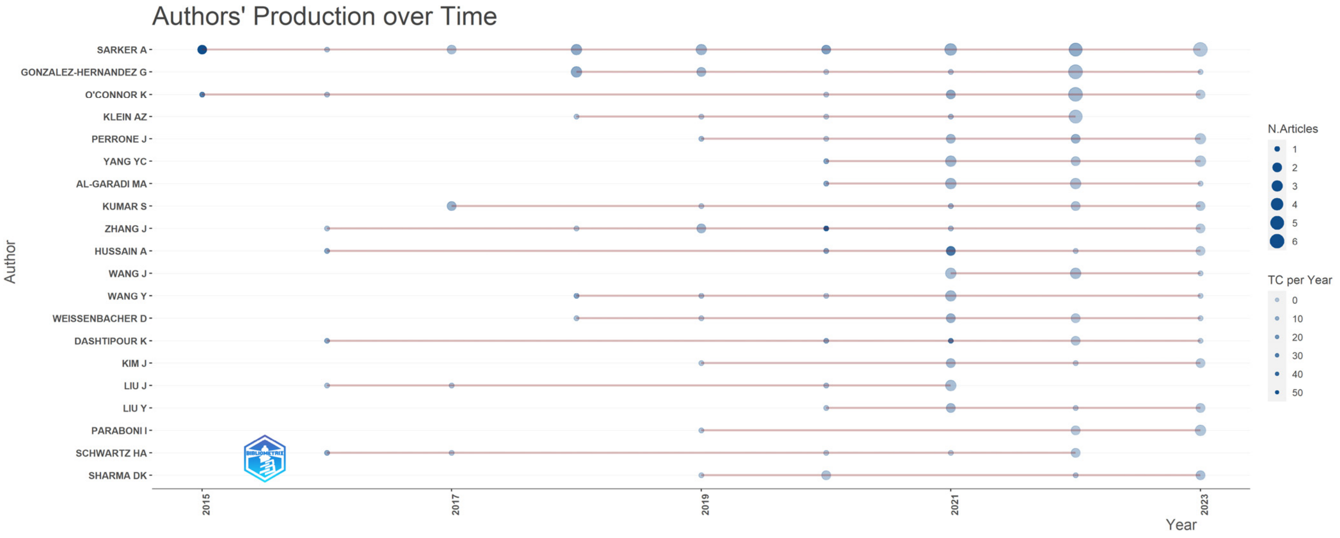

Figure 10. Production of the top 20 authors over time.

The top 21 most relevant affiliations are illustrated in Figure 11. The graph only includes the universities with at least 16 published articles in the area of natural language processing in social media research.

The first position is occupied by the University of Pennsylvania with a remarkable number of 101 papers, followed by Emory University—40 papers, and King Abdulaziz University—38 papers. For the entire list, please consult Figure 11 provided.

Another relevant topic for the bibliometric analysis conducted in this paper is related to the countries of corresponding authors, highlighted in Figure 12.

As expected, the leadership position is held by the USA with an impressive number of papers, namely 402, which registered top values for both single country publications (324 papers) and multiple country publications (78 papers). In second place, China is situated (SCP = 154, MCP = 70), followed closely by India (SCP = 160, MCP = 36). For the entire list, kindly examine Figure 12.

The next figure brings to the fore a visual representation of the worldwide globe, suggestively colored. Therefore, dark shades of blue, such as those found in the case of the USA and China suggest increased scientific production, while at the opposing extreme, shades of grey, such as those in Afghanistan, Costa Rica, and Cuba, indicate a low production of articles published around the topic of natural language processing in social media research. Please see Figure 13.

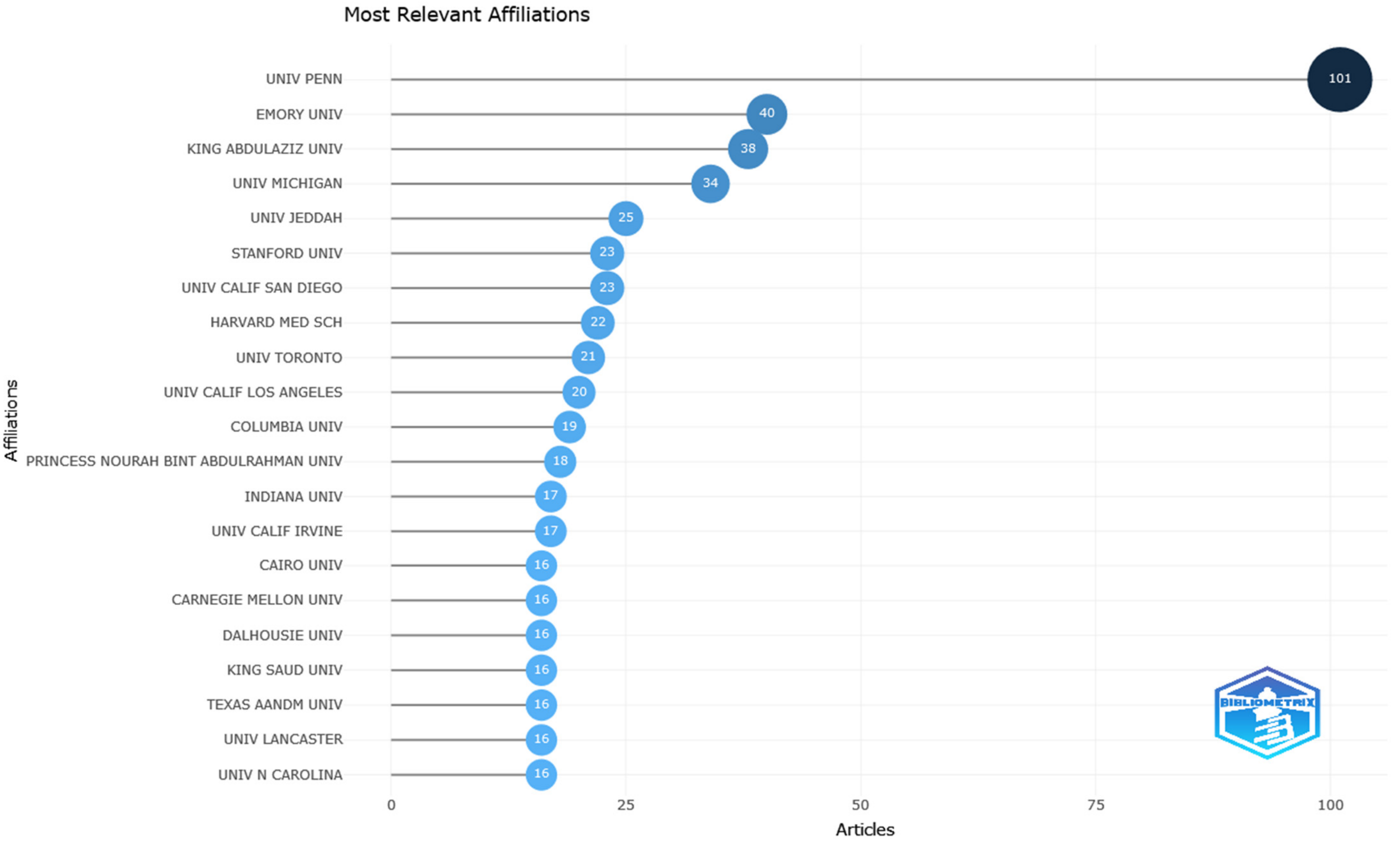

Figure 11. Top 21 most relevant affiliations.

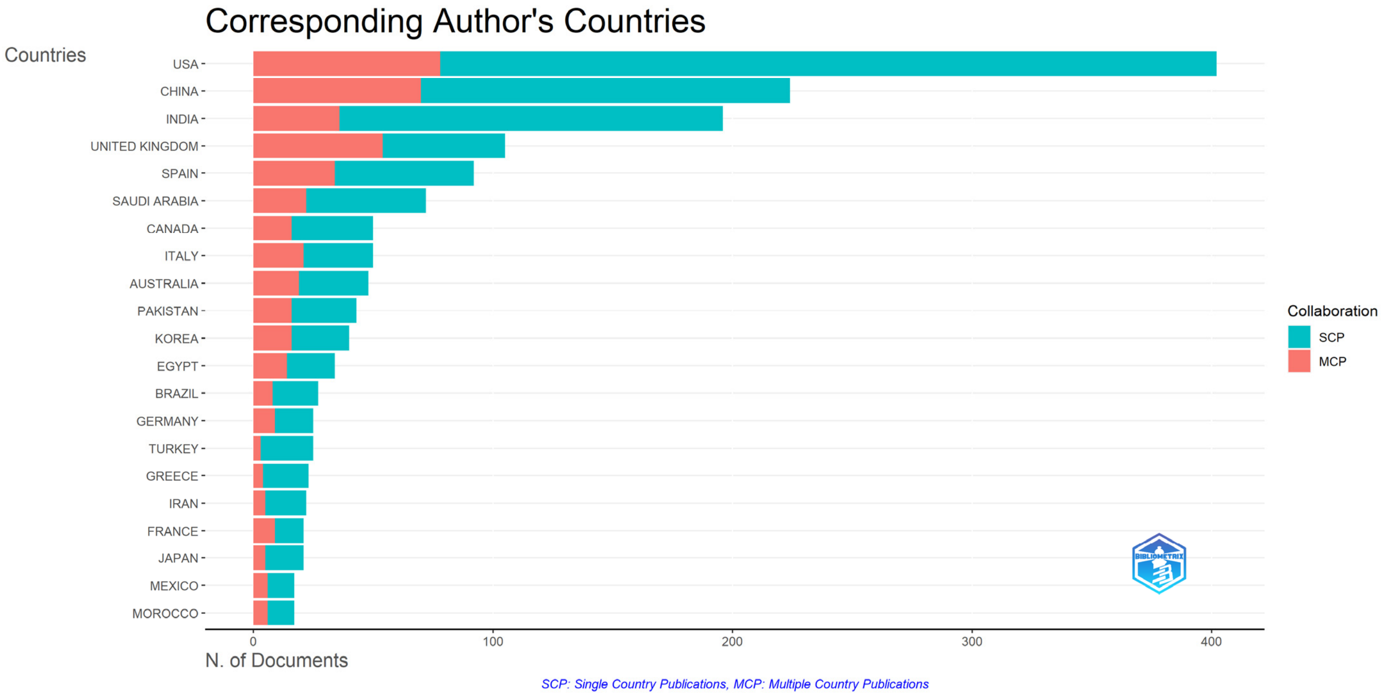

Figure 12. Top 21 most relevant corresponding author's country.

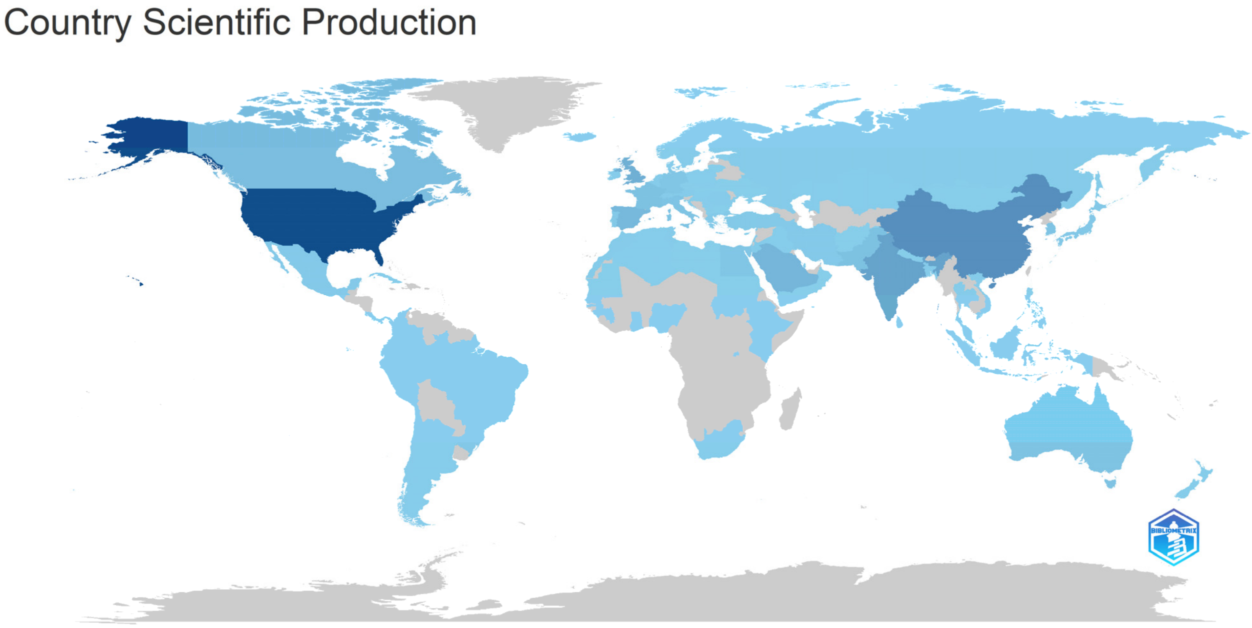

Figure 13. Scientific production based on country.

The most cited countries based on the number of citations are listed in Figure 14.

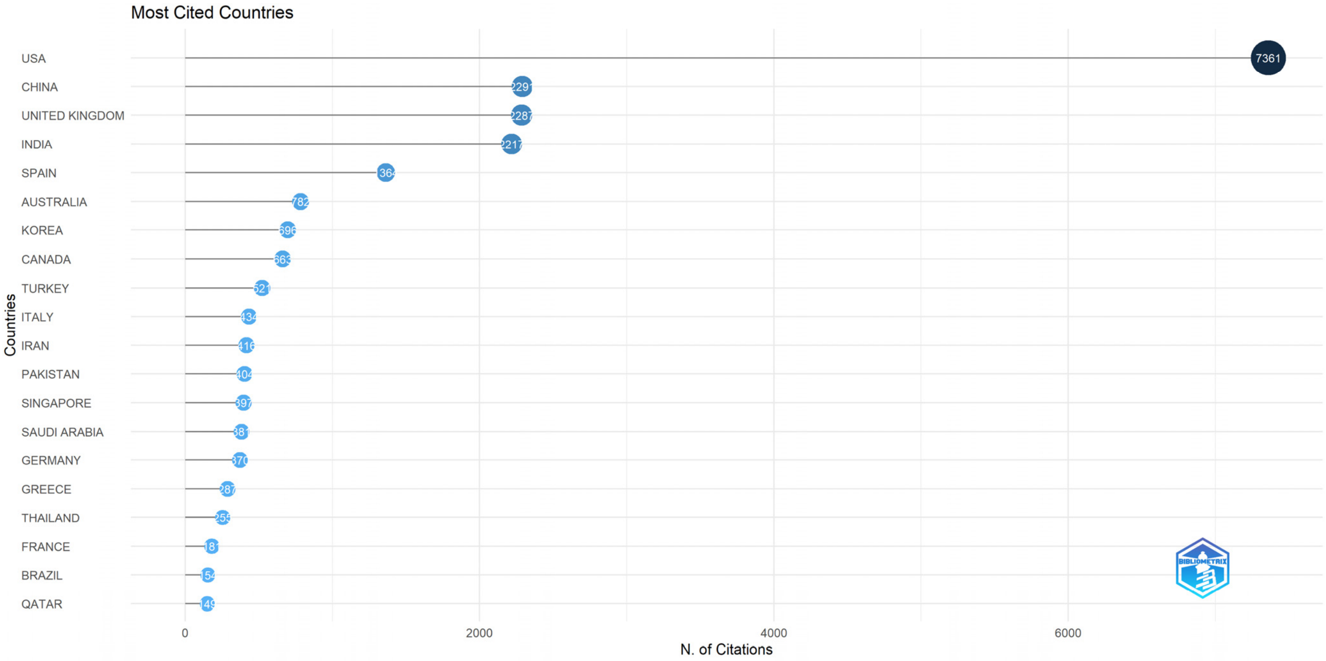

Figure 14. Top 20 countries with the most citations.

A filter criterion was applied when designing the graph, with the intention of excluding countries with less than 149 citations from the top. It was foreseeable that the USA would be found to occupy the first position, with a substantial number of 7361 citations, followed with a considerable difference by China and the United Kingdom, with 2291 and 2287 citations, respectively. See Figure 14 for more information on this matter.

The country collaboration world map is visually exemplified in Figure 15. Of the countries with the greatest number of collaborations, the USA undoubtedly occupies first place, registering associations with no less than 53 countries, including China, the United Kingdom, Canada, and India, and the list continues with numerous others.

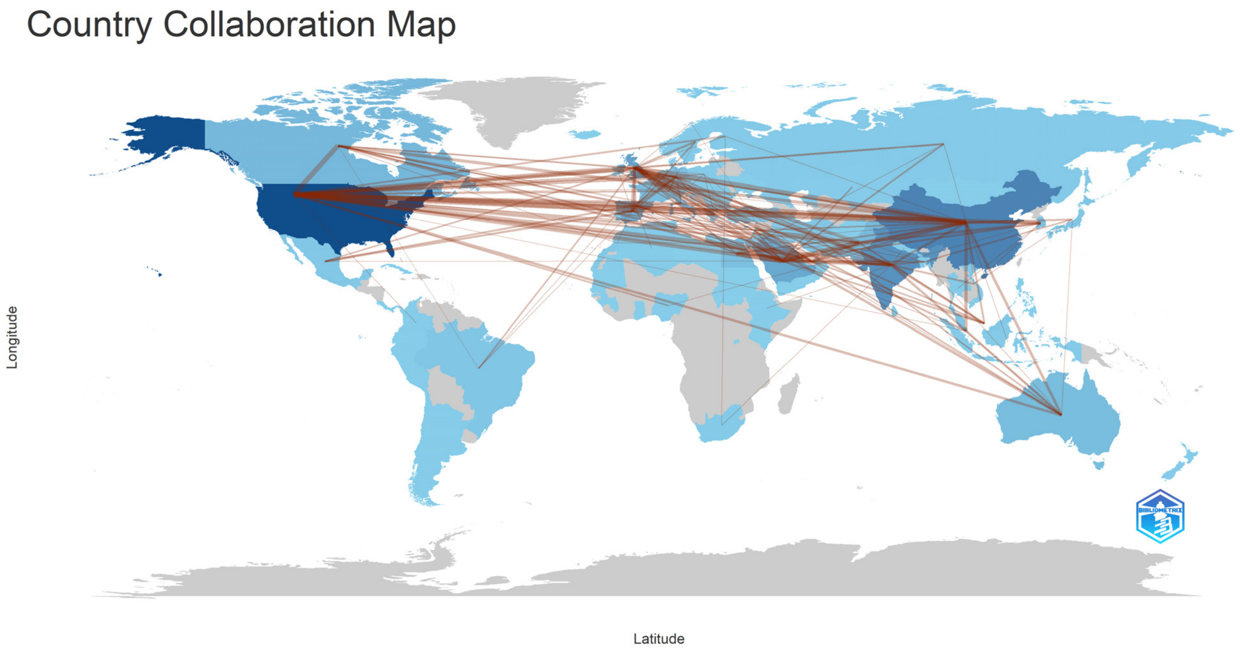

Figure 15. Country collaboration map.

The top 50 authors' collaboration network around the topic of natural language processing in social media research is visually represented with a spectrum of colors in Figure 16.

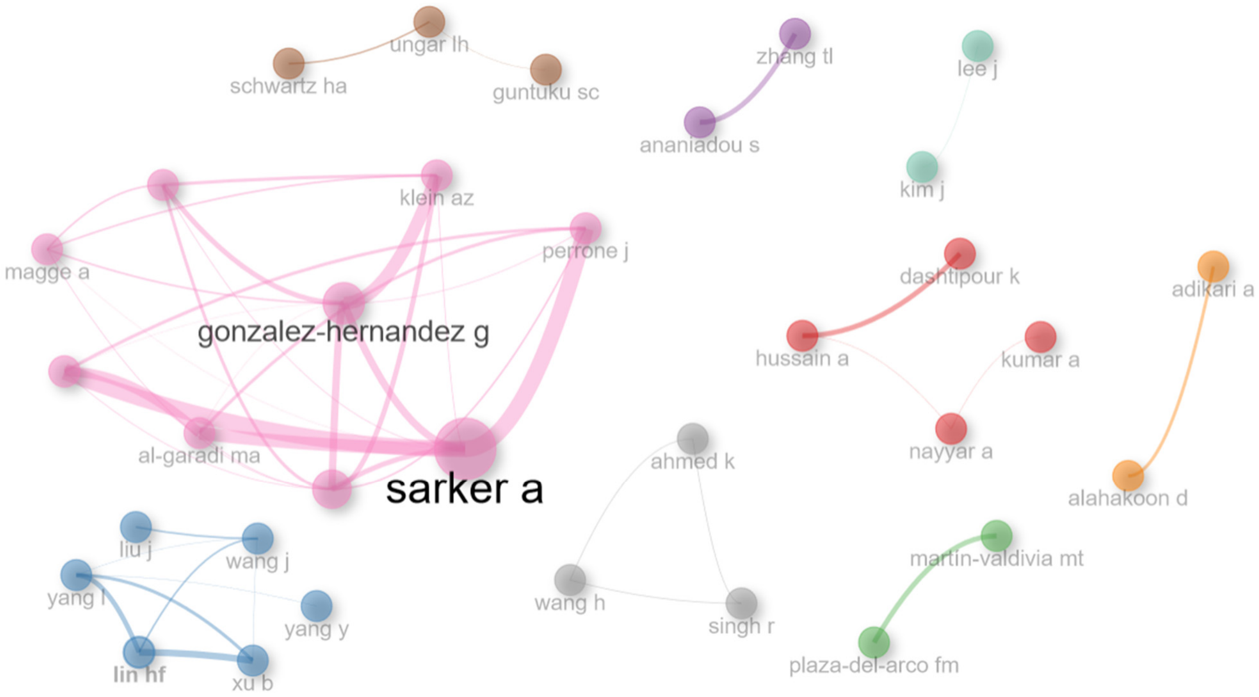

Figure 16. Top 50 authors' collaboration network.

Considering the larger collaboration network—depicted in pink in Figure 16 and further highlighted in grey in Figure 17—it can be observed that it consists of nine authors as follows: Sarker A., Gonzalez-Hernandez G., Al-Garadi M.A., Perrone J., Yang Y.C., O'Connor K., Weissenbacher D., Magge A., and Klein A.Z. In terms of affiliations, it can be observed that the authors belong to various universities such as the University of Pennsylvania—Philadelphia, USA (Sarker A., Gonzalez-Hernandez G., O'Connor K., Perrone J., Weissenbacher D., Klein A.Z.); Emory University—Atlanta, USA (Al-Garadi M.A., Yang Y.C.); and Arizona State University—Scottsdale, USA (Magge A.). As a result, it can be noticed that for the collaboration networks identified, most collaborations occur among authors belonging to the same country. This observation is in line with the information provided in Figure 12, where the number of intra-country collaborations exceeded the number of inter-country collaborations. As a further observation, it can be noted that most of the authors in the above-mentioned cluster have been also listed as the most prominent authors in the dataset (Sarker A., Gonzalez-Hernandez G., O'Connor K., Klein A.Z, Perrone J., Yang Y.C, and Al-Garadi M.A.)

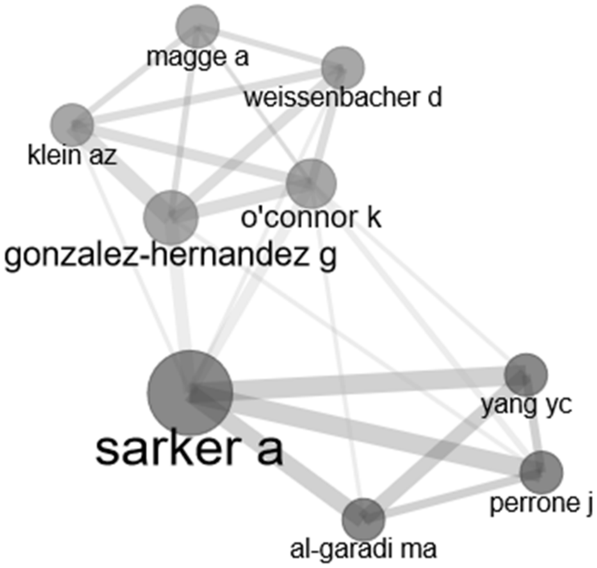

Figure 17. Sarker A. collaboration network.

3.4. Analysis of the Literature

The main goal of this section is to uncover fascinating details associated with the most cited papers included in the data collection set used for the bibliometric analysis conducted in this current research.

Divided into three distinct categories, the analysis of the literature aims to offer readers an overview as comprehensive as possible related to the scientific literature around the area of natural language processing in social media research.

The first sub-section focuses on providing general details about the top 10 most cited papers, including information about the first author, the number of scientists involved in the research and their region of provenance, the source in which the article was published, the year of publication, and the article's reference, including the Digital Object Identifier. In order to delve deeper into the subject, numerical values for important indicators related to the following citations are also provided: total citations (TCs), total citations per year (TCYs), and normalized TCs (NTCs).

The next part offers a brief summary of the content presented in the 10 most globally cited documents by emphasizing the methods used in the analysis, the employed dataset, and the data collection steps. In addition to this, the main purpose of the studies, the goals achieved, and the results obtained are brought to the foreground, with all of this driven by the desire to help readers understand if a specific article is of interest in their individual or professional development.

The final sub-section centers on word analysis, mainly, on the investigation of keywords plus, authors' keywords, bigrams, trigrams, co-occurrence networks, and thematic maps.

3.4.1. Top 10 Most Cited Papers—An Overview

Table 6 sheds light on the main details associated with the top 10 most globally cited documents. At first glance, one can observe that all of the studies were conducted by at least two researchers, a fact that suggests a high degree of collaboration around the topic of natural language processing in fake news research.

In terms of countries, as expected, based on what was already presented previously, the most prevalent place of origin for researchers is the USA. Furthermore, in addition to the studies carried out exclusively by researchers belonging to the same countries, there is also the opportunity to notice articles written by scientists from different regions, which suggests the increased existence of international collaborations around this topic.

Since a hierarchy was established based on the number of total citations, the article written by Ravi et al. [30] is situated in first place, which registered the highest values for two out of three indicators: total citations—675, total citations per year—67.50, and normalized TCs—7.96.

Marked by a considerable difference in terms of the value recorded for total citations indicator, the study conducted by Lee et al. [35] is placed in second position in the ranking, with TCs—389, TCYs—55.57, NTCs—12.08.

The third place is held by the paper belonging to Nikfarjam et al. [36], which logged the following numerical values for the three indicators: TCs—299, TCYs—29.90, NTCs—3.52.

As is discernible by comparing all the values registered in Table 6 for each article, the total citations indicator extends from 210 to 675, TCYs from 20.83 to 67.50, and NTCs from 2.49 to 12.08. These findings provide strong proof that the selected articles are highly relevant within the scientific community.

Table 6. Top 10 most globally cited documents.

No.	Paper (First Author, Year, Journal, Reference)	Number of Authors	Region	Total Citations (TCs)	Total Citations per Year (TCYs)	Normalized TCs (NTCs)
1	Ravi K, 2015, *Knowledge-Based Systems*, [30]	2	India	675	67.50	7.96
2	Lee D, 2018, *Management Science*, [35]	3	USA	389	55.57	12.08
3	Nikfarjam A, 2015, *Journal of the American Medical Informatics Association*, [36]	5	USA	299	29.90	3.52
4	Zubiaga A, 2018, *ACM Computing Surveys*, [37]	5	UK, Germany	274	39.14	8.51
5	Aiello LM, 2013, *IEEE Transactions on Multimedia*, [38]	9	Spain, Greece, UK, France	250	20.83	3.87
6	Middleton SE, 2014, *IEEE Intelligent Systems*, [39]	3	UK	238	21.64	3.89
7	Li LF, 2020, *IEEE Transactions on Computational Social Systems*, [40]	9	China, Hong Kong	232	46.40	10.92
8	Li YY, 2020, *Journal of Marketing Research*, [41]	2	USA	228	45.60	10.73
9	Sarker A, 2015, *Journal of Biomedical Informatics*, [42]	2	USA	211	21.10	2.49
10	Gu YM, 2016, *Transportation Research Part C: Emerging Technologies*, [43]	3	USA	210	23.33	4.78

3.4.2. Top 10 Most Cited Papers—A Review

In the following section, a short review of each paper found in the top 10 most globally cited documents will be provided.

The article that holds the title of the most cited manuscript in the scientific literature around the area of natural language processing in social media research is the one written by Ravi [30], whose relevance and popularity are suggested by the extraordinary number of 675 recorded citations. The article brings to the fore a thorough examination of the papers addressing sentiment analysis, published over a wide period of time—namely 12 years, between 2002 and 2014—emphasizing the methods considered, the datasets used, the techniques of machine learning, and the issues encountered. Therefore, the article is an extremely good guide for researchers who want to have a broad perspective on this domain, along with the matters that require more investigation and potential future work directions.

The research conducted by Lee et al. [35] ranks in the second position at the top, being a comprehensive and excellent investigation, with this hypothesis further supported by the significant amount of citations registered throughout the scientific community. Taking advantage of an extensive dataset of 106,316 messages gathered from 782 distinct companies on a widely accessed online social media platform, Facebook, the analysis aims to comprehend the ways in which various content attributes impact user interactions, illustrated by metrics consisting of likes, comments, shares, and clicks. The study combines natural language processing (NLP) and content analysis to evaluate the textual content of some specific posts, producing 16 binary attributes per message. If one is interested in the results obtained, the paper indicates that people seem to appreciate and interact more with posts that are funny or emotionally oriented and that connect to the personality of the company, rather than merely informative materials, even if, however, certain types of informational material may increase click-through rates. Furthermore, another point addressed in the manuscript is related to the significance of Facebook's EdgeRank algorithm in impacting user engagement and content exposure, accentuating the need for marketers to fully understand and act within the platform's dynamics in order to produce successful content strategies. Long-term brand growth depends on finding a balance between content that addresses the personality of the brand and information material that supports performance marketing goals.

The next article belonging to Nikfarjam et al. [36] is a fascinating manuscript, which, in comparison to the previous ones, is focused on the health domain, more specifically, on the pharmaceutical industry. The researchers propose an innovative artificial-learning-based system referred to as ADRMine, whose primary goal is to extract messages from social networks that mention the adverse reactions experienced by consumers associated with specific drug experiences. The data used for training consist of two corpora: Twitter and DailyStrength (DS), each of them with a set of training and a set of testing, containing messages extracted from online posts. Furthermore, the research employs a variety of techniques, including context, ADR lexicon, parts of speech, and negation identification, while, in order to assist the system for handling unseen tokens, word embeddings and clustering techniques are involved. In terms of the outcomes, the model performs better than anticipated, proving its enhanced effectiveness in collecting messages with respect to the negative effects of drugs from social media platforms, recording considerably better values than basic methods for crucial metrics like precision, recall, and the F-measure.

Rather than focusing on building a new model to identify particular information, the Zubiaga et al. [37] article, in contrast to the other papers presented in this section, provides a thorough overview of the existing literature in the scientific community regarding the detection and resolution of rumors in social networks. The researchers explore rumor classification systems, including their detection and categorization, through in-depth evaluations of numerous approaches across multiple datasets. Furthermore, potential areas for further research are highlighted in addition to the existing challenges. The purpose of the article is not to report experimental results but to give a comprehensive guide for the scientific community by combining and contrasting significant facts from other prior studies.

The next article is written by Aiello et al. [38] and brings to light a different approach of natural language processing in social media research, in an original manner. Starting from six detection algorithms, latent Dirichlet allocation (LDA), document-pivot (Doc-p), graph-based feature-pivot (GFeat-p), frequent pattern mining (FPM), sequential FPM (SFPM), and BNgram, the researchers aim to offer an expanded understanding of each algorithm's performance in the subject of topic detection in messages collected from the Twitter platform, based on precision and recall metrics. Additionally, the process of gathering data is quite thorough, encompassing steps like tokenization, stemming, and aggregation. Thus, the three datasets are divided as follows: 148,652 tweets related to the FA Cup, 474,109 tweets related to Super Tuesday, and 1,247,483 tweets related to the US Elections. The results of the experiment suggest that BNgram performs better, especially in events with a more concentrated theme field, whereas standard topic models (e.g., LDA) function well in events with a tight subjective emphasis.

The development and implementation of a real-time crisis-mapping system is the main focus of the Middleton et al. [39] research. Two sets of data associated with particular disasters in Milan, New York, and Istanbul are collected from the messages posted on social media platform Twitter, and then used for building as well as evaluating the system's efficiency. After performing pre-processing and Treebank word tokenization via the Punkt sentence tokenizer, the location matching algorithm is used. The system's correctness is evaluated and validated using metrics including precision, recall, and F1. The outcome of the investigation demonstrates that the present approach achieves far better values for tweet geoparsing than other solutions currently in use, highlighting the research's significance for the scientific community.

Undoubtedly, the study performed by Li et al. [40] is another highly intriguing article, relevant to our bibliometric analysis around the field of natural language processing in social media research. Its goals include helping the authorities identify the factors that influence users to repost specific information and assisting the government in the development and improvement of strategies related to the publication of information about the COVID-19 pandemic. The dataset consists of postings connected to COVID-19 that were gathered from the Weibo platform using the Weibo API. The posts were then categorized through the use of natural language processing techniques, supervised learning algorithms, and linguistic inquiry and word counting. The study's findings showed that the use of hashtags increased notification reposts; a positive tone amplified the popularity of criticizing opinions; reposts of longer messages happened a greater number of times; and information exposing rumors was more likely to be reposted when it came from people with a wide number of followers. All of the aforementioned findings may provide valuable insights for the authorities and can serve as a foundation in the development of more effective strategies related to the dissemination of information associated with the COVID-19 pandemic.

An intriguing topic—namely, how the presence of an image in a post influences user involvement—is highlighted in the following Li et al. [41] paper, which grabs the reader's interest from the beginning of its contents by being extremely well organized and meticulously detailed. The study focuses on two popular social media platforms, Twitter and Instagram, and the collected datasets are related to US airlines and compact SUV models. Furthermore, three primary effects of picture content are hypothesized: mere presence effect, image characteristics effect, and image–text fit effect. The methods used include propensity score matching and the bivariate zero-inflated negative binomial for Twitter data, while for Instagram, log-linear regression is used. The results of the research reveal that when images are included in postings, users seem to be more involved, demonstrating that factors consisting of the source, quality, and color effects have a direct impact on users' level of engagement. Additionally, the study provides evidence that hyperlinked pictures have negative consequences, suggesting that a substantial number of users are unwilling to allocate extra effort for viewing an image. Lastly, the researchers provide helpful user strategies and underscore the significance of visual content.

In the following research, Sarker et al. [42] desire to enhance the automatic identification of messages on social networks that are directly associated with adverse drug reactions through the use of natural language processing, along with machine learning methods, such as naive Bayes, maximum entropy, and support vector machines. Three datasets collected from Twitter, DailyStrength, and the Adverse Drug Events corpus were used in the investigation. The obtained results demonstrated a considerable improvement in the efficiency of classification, especially with multi-corpus training, and despite discrepancies, merging attributes from several data sources has shown promising outcomes. The paper provides useful insights on how to improve the precision of ADR monitoring on online platforms by integrating pharmacovigilance, machine learning, and natural language processing.

The article by Gu et al. [43] represents the final publication found on the list of the most referenced papers in the field of natural language processing in social media research. The manuscript addresses a novel and captivating subject: enhancing the identification of traffic events in Pittsburgh and Philadelphia in September 2014 using a data collection of tweets collected via the Twitter API. Traffic incident (TI) tweets are identified using a semi-naive Bayes (SNB) classifier, geocoding is utilized to determine the tweets' geographical position, and the supervised latent Dirichlet allocation (sLDA) classifier is used to further refine the incident data. The study's findings reflect the fact that social networks constitute a valuable additional tool for identifying and obtaining information about incidents in real time, and the researchers urge their incorporation in addition to the techniques that are currently in use.

Table 7 provides a brief summary of the content of the above-mentioned documents.

Table 7. A brief summary of the content of the top 10 most globally cited documents.

No.	Paper (First Author, Year, Journal, Reference)	Title	Methods Used	Data	Purpose
1	Ravi K, 2015, *Knowledge-Based Systems*, [30]	A survey on opinion mining and sentiment analysis: Tasks, approaches and applications	Machine Learning-Based Methods, Lexicon-Based Methods, Ontology-Based Methods, Statistical and Mathematical Approaches, Hybrid and Miscellaneous Approaches	Multiple datasets from different online platforms, reviews, social media, etc.	Offer a comprehensive review of the articles written around the area of sentiment analysis, highlighting the main aspects.
2	Lee D, 2018, *Management Science*, [35]	Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook	Content Analysis, Natural Language Processing, Statistical Modeling, Two-Stage Modeling	106,316 messages gathered from Facebook across 782 different companies	Examine the dataset collected from Facebook, so as to understand how content types affect user engagement and offer insights for marketers.
3	Nikfarjam A, 2015, *Journal of the American Medical Informatics Association*, [36]	Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features	Data collection, CRF Features, Learning Word Embeddings, Embedding Cluster Features, Baseline Extraction Techniques	Two corpora: Twitter and DailyStrength (DS)—each of them with a set of training and a set of testing	Build and asses ADRMine, used for extracting the comments from social media posts related to adverse drug reactions.
4	Zubiaga A, 2018, *ACM Computing Surveys*, [37]	Detection and Resolution of Rumours in Social Media: A Survey	Machine learning, Clustering and Heuristics, Feature Analysis, Algorithmic Approaches, Reverse-Tuple-Based Randomized Algorithm, Stance Classification	Many datasets from different platforms, around different subjects	Provide an in-depth overview of the existing literature around the detection and resolution of rumors.

Table 7. *Cont.*

No.	Paper (First Author, Year, Journal, Reference)	Title	Methods Used	Data	Purpose
5	Aiello LM, 2013, *IEEE Transactions on Multimedia*, [38]	Sensing Trending Topics in Twitter	Data Collection and Pre-Processing (Tokenization, Stemming, Aggregation), LDA (Latent Dirichlet Allocation), Doc-p (Document Pivoting), GFeat-p (Graph Feature Pivoting), FPM (Frequent Pattern Mining), SFPM (Sequential FPM), BNgram	Three datasets collected from Twitter: the first one—148,652 tweets, related to the FA Cup; the second one—474,109 tweets, related to Super Tuesday; the third one—1,247,483 tweets, related to the US elections	Asses the performance of the six detection algorithms, using different datasets collected from Twitter.
6	Middleton SE, 2014, *IEEE Intelligent Systems*, [39]	Real-Time Crisis Mapping of Natural Disasters Using Social Media	Data collection and processing, Location-Match Algorithm, Performance Testing, Geoparsing Accuracy Evaluation, Benchmarking, Statistical Analysis of Location Matches	Two Twitter datasets manually marked with locations, streets, and regions	Utilize tweet datasets to develop a real-time crisis-mapping system capable of extracting geographic data for effective catastrophe handling and supervision.
7	Li LF, 2020, *IEEE Transactions on Computational Social Systems*, [40]	Characterizing the Propagation of Situational Information in Social Media During COVID-19 Epidemic: A Case Study on Weibo	Natural Language Processing, Supervised Learning Algorithms (Support Vector Machines, Naive Bayes, Random Forest), Linguistic Inquiry and Word Count	Posts from the Weibo platform, making use of the Weibo API	Provide data-driven insights that may help the authorities improve crisis information by classifying and forecasting the COVID-19-related information on the Weibo platform
8	Li YY, 2020, *Journal of Marketing Research*, [41]	Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement	Propensity Score Matching, Bivariate Zero-Inflated Negative Binomial Model, Log-Linear Regression	Two datasets from Twitter regarding US airlines and compact SUV models; Instagram dataset with image content posts, having at least one like	Examine the way in which image content can influence the popularity of a post from Twitter and Instagram.
9	Sarker A, 2015, *Journal of Biomedical Informatics*, [42]	Portable automatic text classification for adverse drug reaction detection via multi-corpus training	Natural Language Processing, Machine Learning Classifiers	Three datasets collected from Twitter, DailyStrength, and the Adverse Drug Events corpus	Improve the detection of messages related to adverse drug reactions from social media platforms, using machine learning techniques and natural language processing.
10	Gu YM, 2016, *Transportation Research Part C: Emerging Technologies*, [43]	From Twitter to detector: Real-time traffic incident detection using social media data	Data collection and pre-processing, Natural language processing, Supervised Latent Dirichlet Allocation Classifier	Data collected from Twitter (from two regions: the Pittsburgh and Philadelphia Metropolitan Areas) in September 2014	Examine and demonstrate how machine learning techniques may be used to efficiently identify traffic accidents using data collected from Twitter.

3.4.3. Word Analysis

The aim of this section is to provide readers with an in-depth word analysis from multiple perspectives. Therefore, through a comprehensive investigation of keywords plus, authors' keywords, bigrams, trigrams, co-occurrence networks, and thematic maps, a deeper understanding of the topics addressed, areas of interest, subjects encountered, and goals is desired, which will help us draw some conclusions and discover insights into the subject of natural language processing in social media research.

The top 10 most frequent words in keywords plus, found in the data collection set, are extracted in Table 8. At first glance, it can be quickly deduced that most of the articles address topics associated with performing sentiment analysis based on the messages posted on online platforms, mainly Twitter. The key area of interest seems to be health, which is probably a consequence of the famous event, the COVID-19 pandemic, which generated a multitude of debates and fake news intended to induce panic within people in the online environment, a subject that attracted the attention of many scientists.

The most frequent keywords plus are listed here, in accordance with the number of occurrences: "social media"—142 occurrences, "twitter"—125 occurrences, "classification" —83 occurrences, "sentiment analysis"—82 occurrences, "model"—79 occurrences, "information"—68 occurrences, "information"—68 occurrences, "health"—67 occurrences, "impact"—50 occurrences, "media"—47 occurrences, and "online"—36 occurrences.

Table 8. Top 10 most frequent words in keywords plus.

Words	Occurrences
social media	142
twitter	125
classification	83
sentiment analysis	82
model	79
information	68
health	67
impact	50
media	47
online	36

In terms of authors' keywords, the leadership position is held by "natural language processing" with a great number of 969 occurrences, followed by "social media"—609 occurrences, "sentiment analysis"—402 occurrences, "machine learning"—371 occurrences, "deep learning" —272 occurrences, "twitter"—246 occurrences, "COVID-19"—174 occurrences, "text mining"—93 occurrences, "artificial intelligence"—76 occurrences, and "data mining"—72 occurrences. Please see Table 9.

Furthermore, by examining these words, one can easily notice the fact that the premise from which we initially started, as most of the articles extracted during the data collection step address sentiment analysis using messages posted on social media platforms, especially Twitter, in the context of COVID-19 pandemic, is further supported. Under this circumstance, the attention is more focused on the most widespread types of artificial intelligence techniques through which the analyses were carried out.

Table 9. Top 10 most frequent words in authors' keywords.

Words	Occurrences
natural language processing	969
social media	609
sentiment analysis	402
machine learning	371
deep learning	272
twitter	246
COVID-19	174
text mining	93
artificial intelligence	76
data mining	72

Figure 18 represents a colorful representation of the top 50 words found in keywords plus and authors' keywords. The greater the number of occurrences of the word, the greater its size in the image.

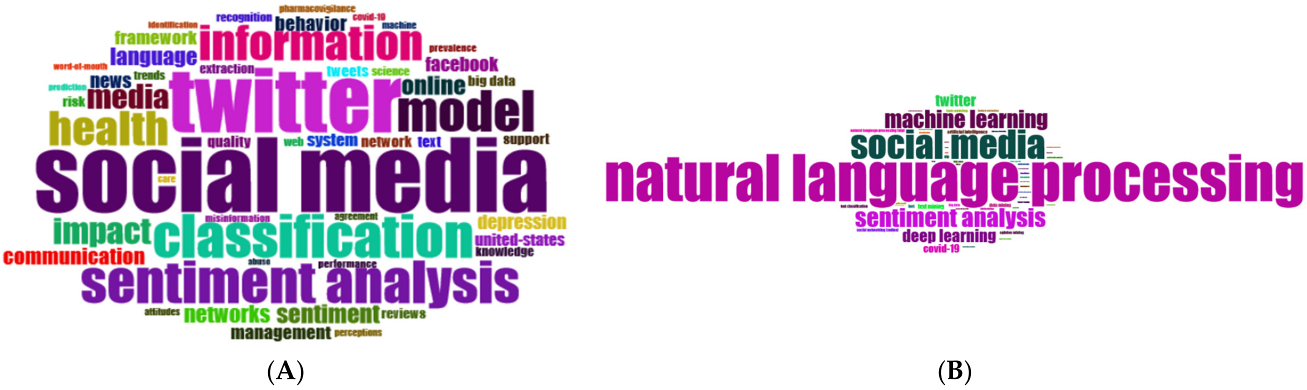

Figure 18. Top 50 words based on keywords plus (**A**) and authors' keywords (**B**).

Table 10 brings to the foreground the top 10 most frequent bigrams that are found in both abstracts and titles throughout the data collection set. Being in line with expectations, the first position is held by "social media", which counted an impressive number of occurrences, 3036 in abstracts and 525 in titles. In second place, for abstracts, "natural language" is situated with 1347 occurrences, while for titles, "sentiment analysis" presents 236 occurrences. Regarding the third position, one can notice "language processing" with 1259 occurrences in abstracts, and "natural language" with 139 occurrences in titles. For more details, kindly refer to Table 10.

Table 10. Top 10 most frequent bigrams in abstracts and titles.

Bigrams in Abstracts	Occurrences	Bigrams in Titles	Occurrences
social media	3036	social media	525
natural language	1347	sentiment analysis	236
language processing	1259	natural language	139
sentiment analysis	845	language processing	134
machine learning	733	deep learning	124
deep learning	440	machine learning	106
processing NLP	377	fake news	62
media platforms	346	media data	56
fake news	285	COVID-19 pandemic	54
media data	274	twitter data	51

Concerning trigrams, the first position, according to the number of occurrences for both abstracts and titles, is held by "natural language processing" (abstracts—1246 occurrences, titles—131 occurrences).

The second and the third place on top are occupied as follows: for bigrams—"natural processing NLP" with 374 occurrences, and "social media platforms" with 344 occurrences, while for trigrams—"social media data" with 55 occurrences and "fake news detection" with 37 occurrences. Please check Table 11.

Figure 19 represents the co-occurrence network for the terms in the authors' keywords. Its main purpose is to shed light on the three delimited clusters, listed as follows:

- Cluster 1, marked in red: natural language processing, social media, sentiment analysis, machine learning, twitter, COVID-19, text mining, artificial intelligence, opinion mining, NLP, topic modeling, big data, infodemiology, mental health, public health, social networks, fake news detection, social media analytics, infoveillance, information extraction, misinformation, reddit, social network analysis, vaccination, tweets, named entity recognition, social media data;

- Cluster 2, marked in blue: deep learning, data mining, natural language processing (NLP), BERT, text classification, fake news, hate speech, LSTM, classification, transfer learning, word embedding, neural networks, social media analysis, text analysis, depression, word2vec;
- Cluster 3, marked in green: social networking (online), feature extraction, blogs, task analysis, transformers, support vector machines.

Table 11. Top 10 most frequent trigrams in abstracts and titles.

Trigrams in Abstracts	Occurrences	Trigrams in Titles	Occurrences
natural language processing	1246	natural language processing	131
language processing NLP	374	social media data	55
social media platforms	344	fake news detection	37
social media data	272	hate speech detection	26
language processing techniques	126	social media text	25
social media posts	125	deep learning model	19
deep learning models	105	social media posts	17
machine learning algorithms	90	social media analysis	13
latent Dirichlet allocation	83	deep learning models	12
bidirectional encoder representations	82	named entity recognition	12

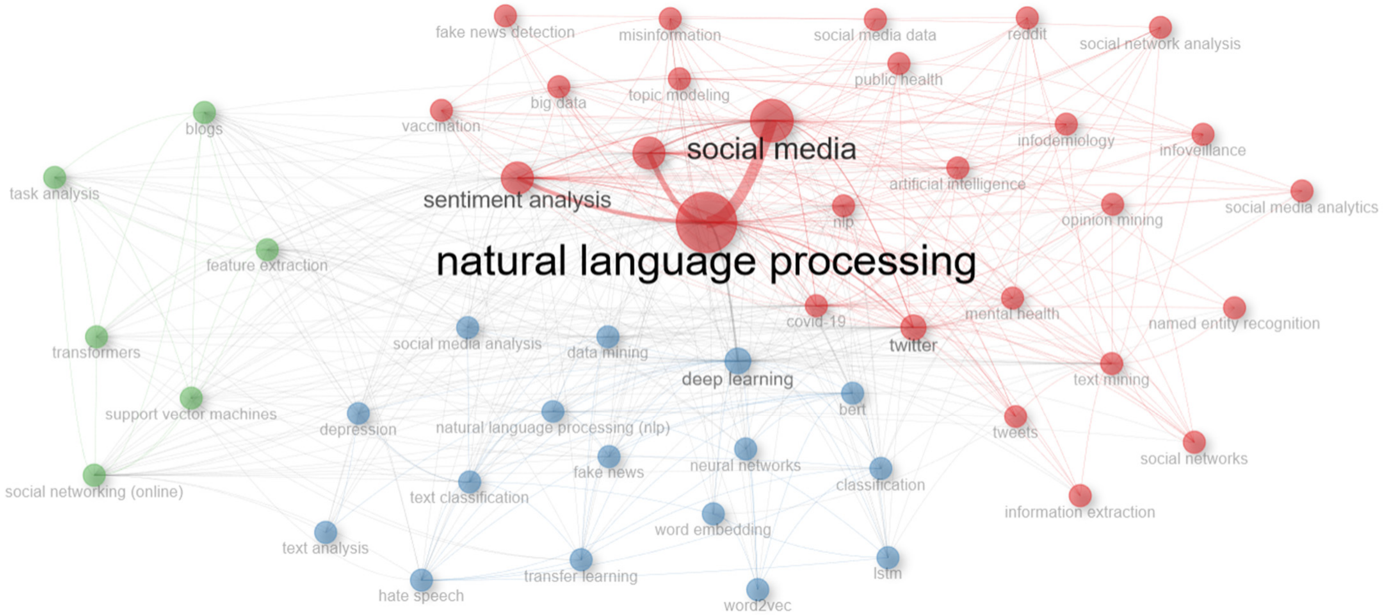

Figure 19. Co-occurrence network for the terms in the authors' keywords.

The thematic map based on the authors' keywords is represented in Figure 20. Based on the visual representation, one can effortlessly note four key themes: niche, motor, basic, and emerging or declining.

The niche themes comprise social media, twitter, and COVID-19, while the motor themes include data mining, natural language processing (NLP), and social networking (online).

Concerning the other two categories, for basic themes, natural language processing, sentiment analysis, and machine learning are identified, whereas on the other hand, for emerging or declining themes, deep learning, BERT, and text classification are depicted.

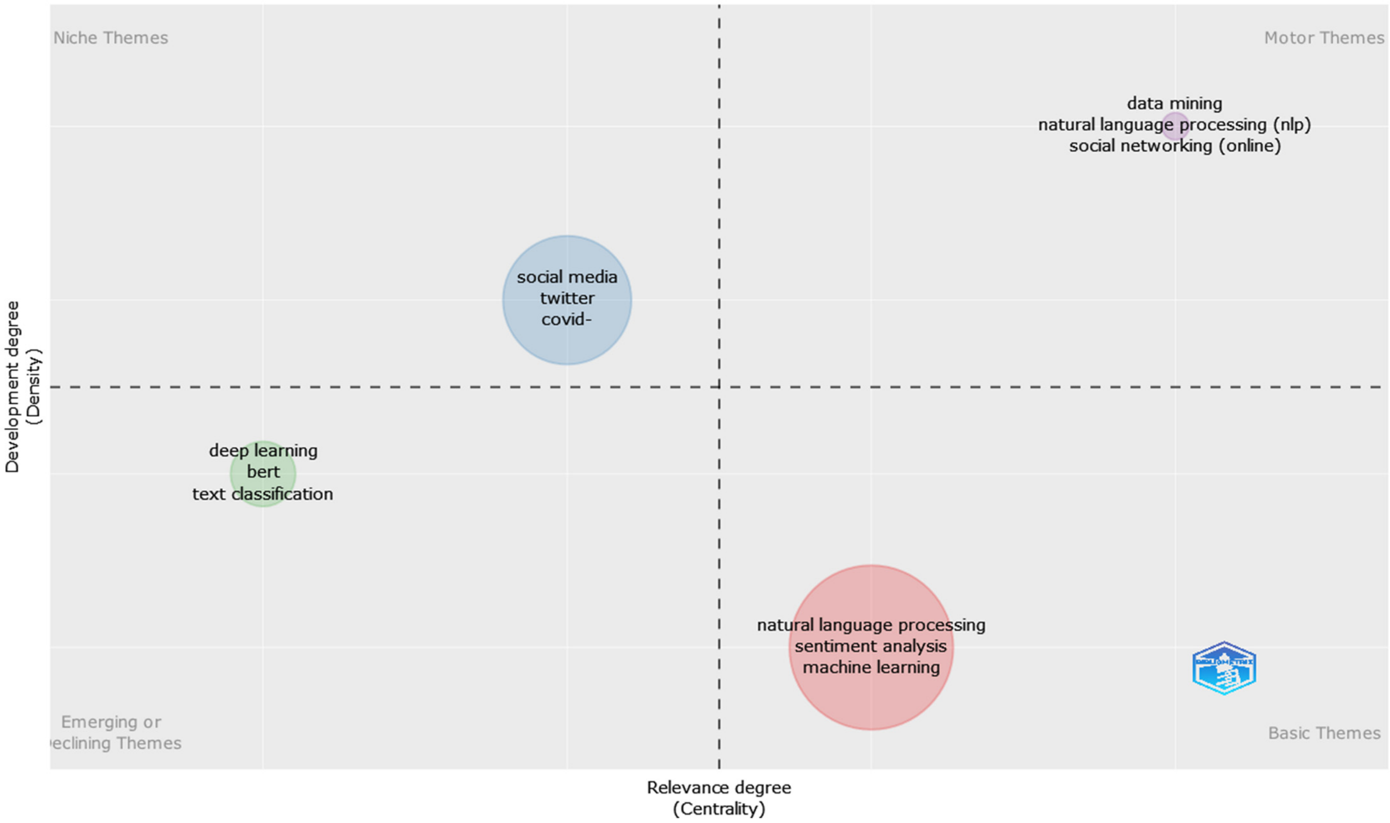

Figure 20. Thematic map based on authors' keywords.

3.5. Mixed Analysis

The ultimate purpose of this chapter is to conduct mixed analyses with the intent of uncovering interesting details and hidden insights about the connection between various perspectives that were brought to the fore above—authors, journals, countries, affiliations, and even keywords.

Figure 21 represents a three-field plot centered on countries (left), authors (middle), and journals (right). By focusing on the first 20 items from each category, some conclusions linked to the connections among the three groups can be drawn.

In terms of authors' countries, it is not surprising based on the information gathered up until this point, that the USA is ranked at the first position, being the main affiliation for most of the scientists who explored the area of natural language processing in social media research in their studies. Regarding the authors, one can easily notice that the leadership position is held by Sarker A, a prolific researcher who contributed to the scientific literature in this area for almost a decade. Moreover, in the figure below, the *Journal of Medical Internet Research* is the most preferred one for publishing articles, and perhaps an explanation for this could be provided by the fact that many of the articles address topics related to the medical area, especially the natural language processing of the messages associated with the COVID-19 pandemic posted online, with this journal being oriented towards this medical domain.

Furthermore, apart from these findings, other noteworthy facets can be considered, such as the following:

- The increased degree of collaboration between researchers belonging to different countries;
- The tendency of multiple scientists to publish their articles in several journals, rather than focusing exclusively on a specific one.

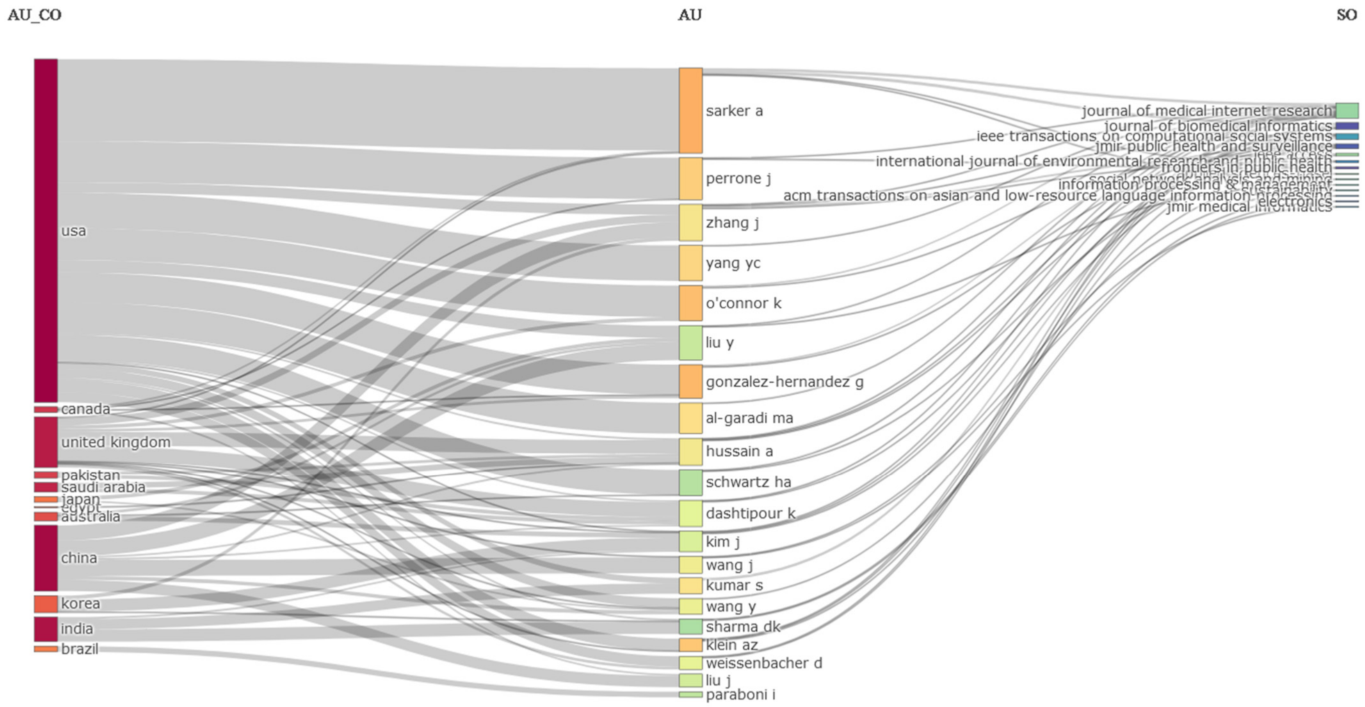

Figure 21. Three-field plot: countries (left), authors (middle), journals (right).

The next three-field plot included in the analysis is oriented toward the interdependencies between affiliations (left), authors (middle), and keywords (right). According to Figure 22, which depicts the first 20 items from each category, the University of Pennsylvania is the most popular affiliation, while Sarker A is the most productive author.

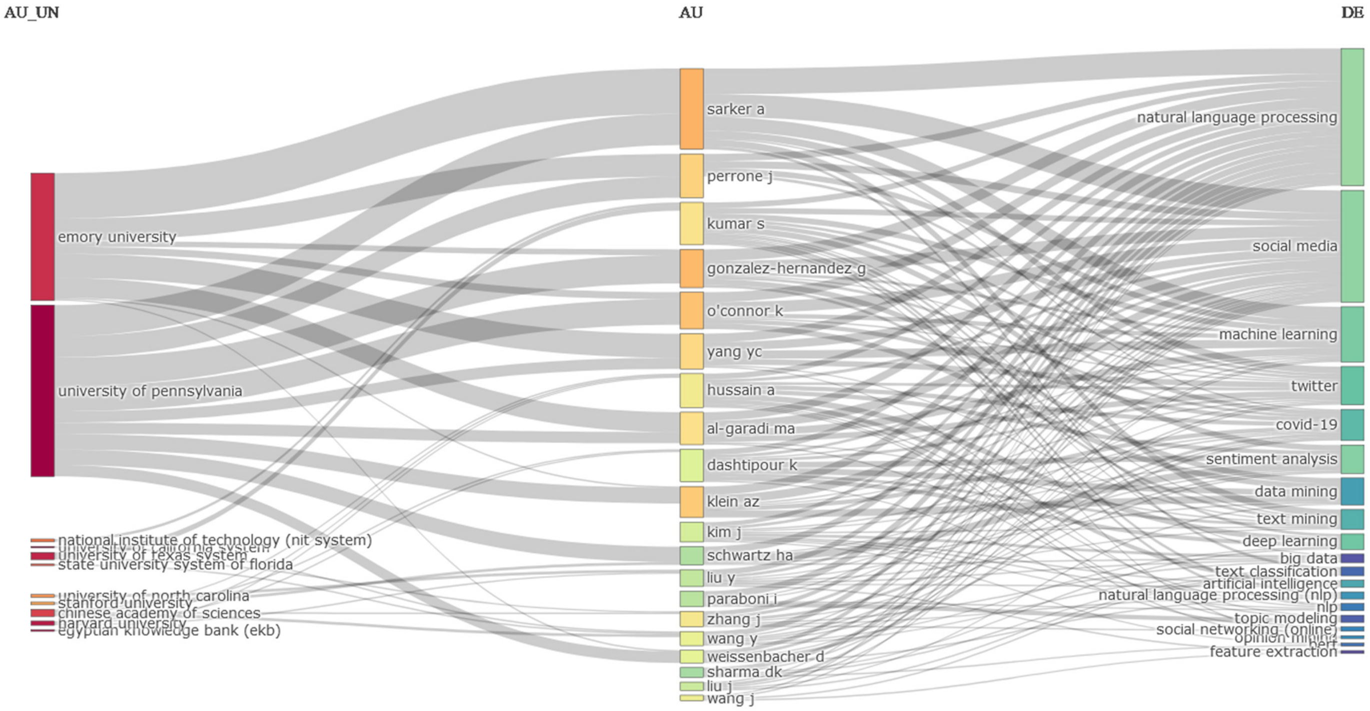

Figure 22. Three-field plot: affiliations (left), authors (middle), keywords (right).

When referring to the keywords, they reinforce all the assumptions from which we initially started. The main area of interest encountered in the vast majority of the articles included in the data collection set is related to the investigation of the online messages posted on Twitter, associated with the COVID-19 pandemic, using techniques like natural language processing, sentiment analysis, data mining, deep learning, and many more.

Moreover, in addition to this, other aspects should be mentioned, including the ones spotted below:

- There are authors who are affiliated with more than one university, and, at the opposite pole, one can notice researchers who are not affiliated with any of the 20 universities included in the top.
- The increased degree of collaboration in this area, together with the affiliation of the authors with universities from different countries, amplify the scientific value of the articles around the area of natural language processing in social media research, benefiting from multiple cultural perspectives and an enriched academic content.

4. Discussion

In view of everything that has been addressed so far, the main objective of the manuscript was to offer a comprehensive bibliometric evaluation around the topic of natural language processing in social media research.

Through the in-depth investigation of multiple perspectives (e.g., references, sources, authors, the existing literature, words, production over time, countries, affiliations), relevant indicators (e.g., collaboration indexes, the h-index), and visual representations (e.g., thematic and collaboration maps, graphs, collaboration networks, co-occurrence networks, three-field plots), the collection of 1852 English articles, published in 2010–2023, was analyzed.

After performing this meticulous assessment, valuable insights and hidden trends were revealed.

There is no doubt that the first starting point of this discussion should definitely be centered on sources. A source is considered truly relevant and influential when obtaining a top position in consecutive bibliometric analyses, conducted on various topics and subjects. This is strong proof for highlighting the source's significant position in the scientific community, its notable presence in several academic articles, by being frequently cited, and the increased interest of the authors to publish their research in it. As for this investigation, the results exposed that the leadership position is held by *IEEE Access*, a popular journal that seems to occupy a standing among additional top ranks, which has been observed in other bibliometric studies throughout the scientific literature [22–24,26,44–48].

In terms of affiliations, the University of Pennsylvania occupies the first position based on the number of papers published around the area of natural language processing in social media research. Its relevance within the scientific community, as well as its exceptional productivity, are further supported by its presence in the top for other bibliometric studies, such as the following ones related to social media research [49], machine learning for mental health in social media [50], machine learning used in cancer research [51], innovation research [52], corporate social responsibility [53], or in general scientific rankings made for specific periods, such as the one conducted by Yan and Zhiping for the 1970–2020 period on the WoS database [54].

Given the countries that are marked as leading contributors in accordance with the number of citations gathered in the area of natural language processing in social media research, one can easily spot that on the top positions are listed as the USA, China, the UK, and India. This is not a surprising outcome, since these regions were also depicted in other bibliometric studies around different subjects, including sentiment analysis, deep learning, health, COVID-19-related topics, and many more. One can refer to the following studies that highlight a similar trend in terms of country contribution as in research by [22–24,46,47].

Coming back to the questions presented in the introduction, in this section, some answers can be outlined, as discussed in the following paragraphs.

Regarding the most cited articles, it was observed that the majority of the studies presented experimental results, while two out of ten were focused on analyzing the existing literature and future work directions in these areas. Apart from this, the methods employed were machine learning techniques, and the data were collected from multiple social network platforms, especially Twitter. The subjects covered varied, including those of the medical domain, rumors, natural disaster crises, traffic incidents, and many more. The number of citations for the top 10 most cited documents fluctuates between 675 and 210.

Regarding the most prolific authors, the foremost position is held by Sarker A with 28 published papers, followed with considerable difference by Gonzalez-Hernandez G, with 14 articles. For both of the authors, the country of origin is the USA, which is also placed in the top for the most cited countries in this domain.

Regarding the collaboration of the authors in the area of natural language processing in social media research, it was observed that the scientists preferred to collaborate with others rather than conduct individual research. The increased degree of collaboration in this area amplifies the scientific value of the articles, benefiting from multiple cultural perspectives and an enriched academic content.

In terms of author production over time, this domain seems to be a topic that has attracted the attention of researchers for quite a long time. With a rather modest beginning, in 2015, only three articles were published (Sarker A—two papers and Gonzalez-Hernandez G—one paper), but from that moment onward, in the succeeding period, a substantial annual increase in the number of publications is observable, which is possibly justified by the occurrence of some events, such as the COVID-19 pandemic or the outbreak of wars, that triggered a multitude of users' reactions on social media.

Considering the analysis of words, one can easily notice that most of the articles address topics associated with performing sentiment analysis based on the messages posted on online platforms, mainly Twitter. The key area of interest seems to be health, probably as a consequence of the famous event, the COVID-19 pandemic, which generated panic throughout the population. When discussing the techniques employed, some of them must be mentioned here: natural language processing, sentiment analysis, machine learning, deep learning, text mining, artificial intelligence, and data mining.

5. Limitations

In line with any scientific work that brings to the fore new discoveries and a valuable contribution to the specialized literature, each manuscript presents certain limitations that must be highlighted objectively, in such a manner that the succeeding researchers who will expand these areas will have the drawbacks in mind and improve them in future studies.

The primary aspect that must be considered is related to the exclusive usage of the WoS database for article selection. The rationale behind our decision was based on several aspects, including its popularity and recognition within the scientific community, its up-to-date version, wide dataset, and its covering of various disciplines. These assumptions were fortified by the fact that most of the bibliometric studies found in the specialized literature, around different subjects, were focused solely in this database [22–27]. This is indeed true, and we fully agree that if other databases had been considered in our analysis, the final data collection set and the results obtained might have differed slightly.

The next limitation that has to be brought to light is associated with the keywords. During the data collection step, the first exclusion criteria applied was focused on only selecting the papers that contained, either in their titles, abstracts, or author's keywords, some specific words related to social media and natural language processing. It may be possible that due to the desire to select only relevant works for this bibliometric study, some papers could have been omitted by mistake, considering certain factors such as linguistic limitations or different terminologies.

The exclusion of non-English papers can also represent an impacting decision on the final outcome of the study. Since this analysis only reflects manuscripts written exclusively in English, the omission of valuable papers written in other languages might have been possible.

Furthermore, another applied exclusion criterion which had major consequences on the reduction in the dataset, more specifically eliminating almost half of the total number of works selected, as spotted in Table 1, is represented by the document type. Considering only the papers marked as articles, some significant studies conducted in books, or even reviews, might have been lost from our analysis.

Last but not least, the final limitation depicted in this manuscript is represented by the predefined time interval. The year 2024 was excluded, but this did not greatly influence the final set of articles, considering the fact that the present bibliometric analysis carried out in this work was conducted at the beginning of 2024.

All the aforementioned exclusion criteria were thoroughly examined and meticulously chosen, guided by the strong desire to include only unique papers relevant to the topic of natural language processing in social media research in our analysis.

Having all these preceding statements pointed out, the purpose of this section was to grant readers a broader vision and understanding behind the article's crafting, highlighting its limitations, and this may represent the foundation of future studies or strategies associated with multiple sectors, such as economics, politics, and even health.

The topic addressed is of utmost importance, thus making it essential to have a comprehensive view of the studies carried out in this field, since the results and insights gained can be valuable weapons in the fight against the dissemination of fake information across social media platforms, ensuring an environment as safe as possible on social networks and also improving the quality of life, along with increasing users' level of happiness in connection with certain events, products, etc.

Regarding future studies, the authors encourage future researchers to delve even deeper into this topic, to include wider datasets, as well as exclude the use of filters as much as possible.

6. Conclusions

In light of everything that has been stated above, the objective of this section is to summarize the main ideas and to bring the article to a unified perspective.

The manuscript's primary intent was to bring to the fore a comprehensive bibliometric analysis, including multiple points of view, on a topic of great relevance nowadays, namely, natural language processing in social media research.

An extensive bibliometric analysis was carried out using the carefully chosen data collection set, which included 1852 relevant publications, gathered by using certain well-established exclusion criteria.

In terms of the most relevant sources, the leadership position is held by *IEEE Access* with an impressive value of 92 published articles in the analyzed area, while the most prolific author is Serker A, with 28 published papers around the subject of natural language processing in social media research.

The most significant affiliation is represented by the University of Pennsylvania with a remarkable number of 101 papers, and regarding the corresponding author's countries, the first place is occupied by the USA.

Reading the insights offered by the word analysis, it can be stated that most of the articles address topics associated with performing sentiment analysis, especially regarding COVID-19-related events, based on the messages posted on online platforms, mainly Twitter. The leading artificial intelligence techniques through which the analyses were carried out include "natural language processing", "machine learning", "deep learning", "text mining", and "data mining".

Apart from these findings, a short summary for each of the top 10 most cited global documents was included, which addressed a topic around natural language processing in social media research, emphasizing the main ideas, the data used, the objectives, and the methods employed. Considering this, some trends may be reinforced here, as follows: the key methods engaged were machine learning techniques (NLP, LDA, vector machines, naive Bayes, etc.), in their investigations, researchers used multiple datasets gathered from various platforms (Twitter, Facebook, DailyStrenght, Weibo, the Adverse Drug Events corpus), and the subjects covered were quite broad, ranging from medical events like the COVID-19 pandemic and drug side effects to topics like marketing, sports, traffic accidents, and many more. Moreover, it was noticed that some articles reported experimental results by building new models and assessing their performance connected to some specific tasks, such as the extraction of relevant information from social media platforms related to a particular matter, while other manuscripts, namely 2 out of 10, were focused on giving a comprehensive guide for the scientific community by combining and contrasting significant facts from other prior studies in the same area.

Apart from these observations, it can be noted that a small portion of the papers deal with political issues (63 papers); most of the papers in this area focused on issues related to political communications, political views of specific countries, political popularity, political marketing, political diffusion, emotions towards candidates during electoral elections, identifying political bots, the refugee crisis, the Ukraine war situation, etc. In terms of political radicalization, the paper by El Barachi et al. [55] distinguishes the use of NLP for analyzing the temporal behavior of extremists on social media, having as a focus point, far-right extremism during the Trump presidency and a number of 259,000 tweets. The authors highlighted that the results obtained through their research are encouraging in the use of advanced social media analytics in the support of effective and timely decision-making [55].

Considering all the observations above and the scientific literature, it can be said that NLP is an extremely powerful tool nowadays, and its integration into the messages posted on social networks is imperative. The increased popularity it enjoys among researchers—and not only them—denotes the accuracy of the results obtained over time. Fake news is a widespread challenge of the technological era in which we live, having many negative implications for humanity. Thus, through its advanced capabilities in continuous evolution, natural language processing is of significant value in the fight against this problem, managing to better identify and filter erroneous, false information with the help of high-performance algorithms. Regarding radical political positions, NLP helps to analyze political speeches, analyze propaganda, and evaluate extremist tendencies, and the list continues with many more. Therefore, NLP is an extremely significant contribution in online discourses, helping researchers to monitor these events and develop optimal strategies for combating these phenomena, as well as increasing people's safety, improving their lifestyle and happiness.

Furthermore, this bibliometric paper outlines certain types of analysis associated with natural language processing, such as n-gram analysis, through the lens of reviewing the most frequently cited articles published in the present area of interest. This review's key objectives were to introduce the reader to the subject matter, show an overview that comprises the present literature, and highlight the subjects that previous researchers have addressed in order to help an audience understand whether or not the specific articles are of interest to them. For a deeper analysis, a rigorous study of the most cited articles is also recommended, and a potential direction for continuing the manuscript could, undoubtedly, explain the significance of each type of analysis for social media research.

As mentioned in the introduction, the purpose of this article was to present a bibliometric analysis as comprehensively as possible, covering various spheres, such as authors, sources, existing works in the specialized literature, word analysis, mixed analysis, and many others. Through the obtained results, this manuscript brings to the fore the major themes addressed by researchers interested in this area, as well as current trends together with valuable information, and serves as a powerful tool for future studies. Although it does not directly explore the limitations and challenges of NLP in social media research, this paper can be considered an influential resource, providing a solid basis for understanding the limitations and challenges of NLP, current technological issues, and potential directions.

To conclude, natural language processing in social media research is not only a subject approached for reasons of curiosity, but it is a matter with considerable far-reaching effects and implications in society. The accelerated evolution of technology, together with advanced machine learning techniques, have extraordinary power to analyze texts and extract relevant information.

As a result, a thorough analysis of the messages sent on social networks using natural language processing can reveal a wealth of crucial information that may be further used to develop strategies aimed at strengthening online security, enhancing user happiness on social media platforms, identifying the problems that society is currently facing, preventing the spread of false and inaccurate information meant to induce panic within society, and improving political and economic decisions.

Author Contributions: Conceptualization, A.S. (Andra Sandu) and C.D.; Data curation, A.S. (Andra Sandu), A.S. (Aurelia Stănescu) and C.D.; Formal analysis, A.S. (Andra Sandu), L.-A.C., A.S. (Aurelia Stănescu) and C.D.; Investigation, A.S. (Andra Sandu), L.-A.C., A.S. (Aurelia Stănescu) and C.D.; Methodology, A.S. (Andra Sandu), L.-A.C. and C.D.; Software, A.S. (Andra Sandu), L.-A.C. and C.D.; Supervision, L.-A.C. and C.D.; Validation, A.S. (Andra Sandu), L.-A.C., A.S. (Aurelia Stănescu) and C.D.; Visualization, A.S. (Andra Sandu), A.S. (Aurelia Stănescu) and C.D.; Writing—original draft, A.S. (Andra Sandu); Writing—review and editing, L.-A.C., A.S. (Aurelia Stănescu) and C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Deans, P.C.; Tretola, B.J.M. The Evolution of Social Media and Its Impact on Organizations and Leaders. *J. Organ. Comput. Electron. Commer.* **2018**, *28*, 173–192. [CrossRef]
2. Edosomwan, S.; Prakasan, S.K.; Kouame, D.; Watson, J.; Seymour, T. The History of Social Media and Its Impact on Business. *J. Appl. Manag. Entrep.* **2011**, *16*, 79–91.
3. Sharma, S.; Verma, H.V. Social Media Marketing: Evolution and Change. *Soc. Media Mark. Evol. Chang.* **2018**, *25*, 19–36. [CrossRef]
4. Callier, V. Machine Learning in Evolutionary Studies Comes of Age. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2205058119. [CrossRef] [PubMed]
5. Ljubobratović, D.; Vuković, M.; Brkić Bakarić, M.; Jemrić, T.; Matetić, M. Assessment of Various Machine Learning Models for Peach Maturity Prediction Using Non-Destructive Sensor Data. *Sensors* **2022**, *22*, 5791. [CrossRef] [PubMed]
6. Telikani, A.; Tahmassebi, A.; Banzhaf, W.; Gandomi, A.H. Evolutionary Machine Learning: A Survey. *ACM Comput. Surv.* **2021**, *54*, 1–35. [CrossRef]
7. Ortiz-Garces, I.; Govea, J.; Andrade, R.O.; Villegas-Ch, W. Optimizing Chatbot Effectiveness through Advanced Syntactic Analysis: A Comprehensive Study in Natural Language Processing. *Appl. Sci.* **2024**, *14*, 1737. [CrossRef]
8. Chang, K.-H. Natural Language Processing: Recent Development and Applications. *Appl. Sci.* **2023**, *13*, 11395. [CrossRef]
9. Hirschberg, J.; Manning, C.D. Advances in Natural Language Processing. *Science* **2015**, *349*, 261–266. [CrossRef] [PubMed]
10. Zhang, D.D.; Wang, J.; Sun, M. The Progress That Natural Language Processing Has Made Towards Human-Level AI. *ResearchGate* **2020**, *3*, 38–47. [CrossRef]
11. Jiang, Y.; Pang, P.C.-I.; Wong, D.; Kan, H.Y. Natural Language Processing Adoption in Governments and Future Research Directions: A Systematic Review. *Appl. Sci.* **2023**, *13*, 12346. [CrossRef]
12. Pandey, K.K.; Thorat, M.; Joshi, A.; D, S.; Hussein, A.; Alazzam, M.B. Natural Language Processing for Sentiment Analysis in Social Media Marketing. In Proceedings of the 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 12–13 May 2023; pp. 326–330. [CrossRef]
13. Al-Saif, H.F.; Al-Dossari, H.Z. Exploring the Role of Emotions in Arabic Rumor Detection in Social Media. *Appl. Sci.* **2023**, *13*, 8815. [CrossRef]
14. Boon-Itt, S.; Skunkan, Y. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *JMIR Public Health Surveill.* **2020**, *6*, e21978. [CrossRef] [PubMed]
15. Block, J.H.; Fisch, C. Eight Tips and Questions for Your Bibliographic Study in Business and Management Research. *Manag. Rev. Q.* **2020**, *70*, 307–312. [CrossRef]
16. WoS Web of Science. Available online: https://www.webofknowledge.com (accessed on 9 September 2023).
17. Liu, W. The Data Source of This Study Is Web of Science Core Collection? Not Enough. *Scientometrics* **2019**, *121*, 1815–1824. [CrossRef]

18. Liu, F. Retrieval Strategy and Possible Explanations for the Abnormal Growth of Research Publications: Re-Evaluating a Bibliometric Analysis of Climate Change. *Scientometrics* **2022**, *128*, 853–859. [CrossRef] [PubMed]
19. Bakır, M.; Özdemir, E.; Akan, Ş.; Atalık, Ö. A Bibliometric Analysis of Airport Service Quality. *J. Air Transp. Manag.* **2022**, *104*, 102273. [CrossRef]
20. Cobo, M.J.; Martínez, M.A.; Gutiérrez-Salcedo, M.; Fujita, H.; Herrera-Viedma, E. 25 Years at Knowledge-Based Systems: A Bibliometric Analysis. *Knowl.-Based Syst.* **2015**, *80*, 3–13. [CrossRef]
21. Mulet-Forteza, C.; Martorell-Cunill, O.; Merigó, J.M.; Genovart-Balaguer, J.; Mauleon-Mendez, E. Twenty Five Years of the Journal of Travel & Tourism Marketing: A Bibliometric Ranking. *J. Travel Tour. Mark.* **2018**, *35*, 1201–1221. [CrossRef]
22. Sandu, A.; Cotfas, L.-A.; Delcea, C.; Crăciun, L.; Molanescu, A.G. Sentiment Analysis in the Age of COVID-19: A Bibliometric Perspective. *Information* **2023**, *14*, 659. [CrossRef]
23. Sandu, A.; Ioanăș, I.; Delcea, C.; Florescu, M.-S.; Cotfas, L.-A. Numbers Do Not Lie: A Bibliometric Examination of Machine Learning Techniques in Fake News Research. *Algorithms* **2024**, *17*, 70. [CrossRef]
24. Sandu, A.; Ioanăș, I.; Delcea, C.; Geantă, L.-M.; Cotfas, L.-A. Mapping the Landscape of Misinformation Detection: A Bibliometric Approach. *Information* **2024**, *15*, 60. [CrossRef]
25. Delcea, C. Grey Systems Theory in Economics—Bibliometric Analysis and Applications' Overview. *Grey Syst. Theory Appl.* **2015**, *5*, 244–262. [CrossRef]
26. Delcea, C.; Domenteanu, A.; Ioanăș, C.; Vargas, V.M.; Ciucu-Durnoi, A.N. Quantifying Neutrosophic Research: A Bibliometric Study. *Axioms* **2023**, *12*, 1083. [CrossRef]
27. Domenteanu, A.; Delcea, C.; Chirita, N.; Ioanăș, C. From Data to Insights: A Bibliometric Assessment of Agent-Based Modeling Applications in Transportation. *Appl. Sci.* **2023**, *13*, 12693. [CrossRef]
28. WoS Document Types. Available online: https://webofscience.help.clarivate.com/en-us/Content/document-types.html (accessed on 3 December 2023).
29. Aria, M.; Cuccurullo, C. Bibliometrix: An R-Tool for Comprehensive Science Mapping Analysis. *J. Informetr.* **2017**, *11*, 959–975. [CrossRef]
30. Ravi, K.; Ravi, V. A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications. *Knowl.-Based Syst.* **2015**, *89*, 14–46. [CrossRef]
31. Reyes, A.; Rosso, P.; Buscaldi, D. From Humor Recognition to Irony Detection: The Figurative Language of Social Media. *Data Knowl. Eng.* **2012**, *74*, 1–12. [CrossRef]
32. Montoyo, A.; Martínez-Barco, P.; Balahur, A. Subjectivity and Sentiment Analysis: An Overview of the Current State of the Area and Envisaged Developments. *Decis. Support Syst.* **2012**, *53*, 675–679. [CrossRef]
33. Wardikar, V. Application of Bradford's Law of Scattering to the Literature of Library & Information Science: A Study of Doctoral Theses Citations Submitted to the Universities of Maharashtra, India. *Libr. Philos. Pract.* **2013**, *15*, 1–45.
34. RDRR Website Bradford: Bradford's Law in Bibliometrix: Comprehensive Science Mapping Analysis. Available online: https://rdrr.io/cran/bibliometrix/man/bradford.html (accessed on 21 November 2023).
35. Lee, D.; Hosanagar, K.; Nair, H.S. Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. *Manag. Sci.* **2018**, *64*, 4967–5460. [CrossRef]
36. Nikfarjam, A.; Sarker, A.; O'Connor, K.; Ginn, R.; Gonzalez, G. Pharmacovigilance from Social Media: Mining Adverse Drug Reaction Mentions Using Sequence Labeling with Word Embedding Cluster Features. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 671–681. [CrossRef] [PubMed]
37. Zubiaga, A.; Aker, A.; Bontcheva, K.; Liakata, M.; Procter, R. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
38. Aiello, L.M.; Petkos, G.; Martin, C.; Corney, D.; Papadopoulos, S.; Skraba, R.; Göker, A.; Kompatsiaris, I.; Jaimes, A. Sensing Trending Topics in Twitter. *IEEE Trans. Multimed.* **2013**, *15*, 1268–1282. [CrossRef]
39. Middleton, S.E.; Middleton, L.; Modafferi, S. Real-Time Crisis Mapping of Natural Disasters Using Social Media. *IEEE Intell. Syst.* **2023**, *29*, 9–17. [CrossRef]
40. Li, L.; Zhang, Q.; Wang, X.; Zhang, J.; Wang, T.; Gao, T.-L.; Duan, W.; Tsoi, K.K.-F.; Wang, F.-Y. Characterizing the Propagation of Situational Information in Social Media During COVID-19 Epidemic: A Case Study on Weibo. *IEEE Trans. Comput. Soc. Syst.* **2020**, *7*, 556–562. [CrossRef]
41. Li, Y.; Xie, Y. Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement. *J. Mark. Res.* **2019**, *57*, 1–9. [CrossRef]
42. Sarker, A.; Gonzalez, G. Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-Corpus Training. *J. Biomed. Inform.* **2015**, *53*, 196–207. [CrossRef] [PubMed]
43. Gu, Y.; Qian, Z.; Chen, F. From Twitter to Detector: Real-Time Traffic Incident Detection Using Social Media Data. *Transp. Res. Part C Emerg. Technol.* **2016**, *67*, 321–342. [CrossRef]
44. Puteh, N. Sentiment Analysis with Deep Learning: A Bibliometric Review. *Turk. J. Comput. Math. Educ.* **2021**, *12*, 1509–1519.
45. Sarirete, A. A Bibliometric Analysis of COVID-19 Vaccines and Sentiment Analysis. *Procedia Comput. Sci.* **2021**, *194*, 280–287. [CrossRef] [PubMed]
46. Michailidis, P. Visualizing Social Media Research in the Age of COVID-19. *Information* **2022**, *13*, 372. [CrossRef]

47. Mahajan, R.; Gupta, P. A Bibliometric Analysis on the Dissemination of COVID-19 Vaccine Misinformation on Social Media. *J. Content Community Commun.* **2021**, *14*, 218–229. [CrossRef]
48. Arora, S.; Majumdar, A. Machine Learning and Soft Computing Applications in Textile and Clothing Supply Chain: Bibliometric and Network Analyses to Delineate Future Research Agenda. *Expert Syst. Appl.* **2022**, *200*, 117000. [CrossRef]
49. Gan, C.; Wang, W. A Bibliometric Analysis of Social Media Research from the Perspective of Library and Information Science. In *Digital Services and Information Intelligence*; Li, H., Mäntymäki, M., Zhang, X., Eds.; IFIP Advances in Information and Communication Technology; Springer: Berlin/Heidelberg, Germany, 2014; Volume 445, pp. 23–32, ISBN 978-3-662-45525-8.
50. Kim, J.; Lee, D.; Park, E. Machine Learning for Mental Health in Social Media: Bibliometric Study. *J. Med. Internet Res.* **2021**, *23*, e24870. [CrossRef] [PubMed]
51. Lin, L.; Liang, L.; Wang, M.; Huang, R.; Gong, M.; Song, G.; Hao, T. A Bibliometric Analysis of Worldwide Cancer Research Using Machine Learning Methods. *Cancer Innov.* **2023**, *2*, 219–232. [CrossRef] [PubMed]
52. Cancino, C.A.; Merigó, J.M.; Coronado, F.C. A Bibliometric Analysis of Leading Universities in Innovation Research. *J. Innov. Knowl.* **2017**, *2*, 106–124. [CrossRef]
53. Cucari, N.; Tutore, I.; Montera, R.; Profita, S. A Bibliometric Performance Analysis of Publication Productivity in the Corporate Social Responsibility Field: Outcomes of SciVal Analytics. *Corp. Soc. Responsib. Environ. Manag.* **2023**, *30*, 1–16. [CrossRef]
54. Yan, L.; Zhiping, W. Mapping the Literature on Academic Publishing: A Bibliometric Analysis on WOS. *Sage Open* **2023**, *13*, 21582440231158562. [CrossRef]
55. El Barachi, M.; Mathew, S.S.; Oroumchian, F.; Ajala, I.; Lutfi, S.; Yasin, R. Leveraging Natural Language Processing to Analyse the Temporal Behavior of Extremists on Social Media. *J. Commun. Softw. Syst.* **2022**, *18*, 195–207. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.