SCIENTIFIC Publishing Limited

Digital Technologies Research and Applications

http://ojs.ukscip.com/index.php/dtra

Article

Mapping the Intersection of Artificial Intelligence and Sociolinguistics: A Bibliometric and Keyword-Based Content Analysis

Rugaiyah [1,*], Andi Idayani [1], Roziah [2], Nunuk Suryanti [3] and Novri Gazali [4]

[1] Department of English Education, Faculty of Teacher Training and Education, Universitas Islam Riau, Pekanbaru 28125, Indonesia

[2] Department of Indonesian Language and Literature Education, Faculty of Teacher Training and Education, Universitas Islam Riau, Pekanbaru 28125, Indonesia

[3] Department of Accounting Education, Faculty of Teacher Training and Education, Universitas Islam Riau, Pekanbaru 28125, Indonesia

[4] Department of Physical Education, Faculty of Teacher Training and Education, Universitas Islam Riau, Pekanbaru 28125, Indonesia

[*] Correspondence: ruqaiyah@edu.uir.ac.id

Received: 14 January 2026; **Revised:** 3 February 2026; **Accepted:** 3 March 2026; **Published:** 13 April 2026

Abstract: This research investigates the dynamic relationship of Artificial Intelligence (AI) and Sociolinguistics through bibliometric mapping in association with keyword content analysis. Utilizing 69 extracted publications (2013–2024) after systematic deduplication, the study combines quantitative trend analysis with keyword-based thematic interpretation. From an initial collection of 98 records obtained from Scopus (n = 64) and Web of Science (n = 34), a subset of 48 publications was sampled further pursuant to their conceptual relevance. Bibliometric analysis with the software ScientoPy and VOSviewer was employed to reveal publication trajectories, top contributors, influential journals, geographic patterns, and knowledge hot spots. This mapping was supplemented with a qualitative examination of the space mapped using five major terms: Computational Sociolinguistics, Natural Language Processing (NLP), ChatGPT, language and machine learning enabling us to track prevalent themes and concepts structuring the field. These results indicate that scholarly interest in the sociolinguistic aspects of AI-mediated communication has grown substantially, especially pertaining to language ideology, identity construction, and algorithmic influence on discourse. Instead of portraying computational methods as passive and neutral tools, the findings imply that technology such as NLP and large language models can be seen as both reproducing and destabilizing linguistic hierarchies, bringing to light critical questions regarding representation, diversity, and equity in digital space. In this work, we map the intersection of AI and Sociolinguistics through a combination of bibliometric mapping and keyword-based interpretation, thus giving an overview of how the field has evolved over time. This finding implies that debates about ethical and culturally inclusive AI design are coalescing into prominence in the literature.

Keywords: Artificial Intelligence; Sociolinguistics Ideology; Language Standarization; Algorithmic Mediation; Identity Formation; ChatGPT

1. Introduction

Artificial Intelligence (AI) is not just transforming technology and industry, but also altering the way users communicate and express themselves in modern communication systems. This mechanism has recast AI as a strategic instrument of ideological production via reproduction of social values in language. It has also been found in related researches that AI does have bias in NLP, such as dialect difference [1]. Building on this concern, Sheng et al. [2] and Blasi et al. [3] point out that while AI is not linguistically neutral, it simply reinforces, if anything, dominant language representations, and, by extension, perhaps marginalizes those variation types that were simply not part of the main training example.

From a language ideology perspective, AI can contribute to supporting linguistic hegemony, and especially the privilege of digitally powerful languages like English, a situation that explicitly or implicitly contributes to the marginalization of minority languages [4,5]. This inequality represents a new instance of linguistic inequality in the globalized digital sphere [6]. In the daily interactions, AI users (chatbots, virtual assistants) are exposed to systems that reflect a form of meta-linguistic normalization, reinforcing the dominant linguistic practices and discarding sociolinguistic diversity [7,8].

AI transforms portrayals of everyday language (in public discourse as well as in informal intersubjective communication). Text-based NLP assists in shaping language patterns online [9] while AI helps influence pronunciation, intonation, and even users' sense of linguistic identity. In subsequent work, Twitter has served as an important site for discourse analysis of digital sociolinguistic dynamics, including: syntactic change in dialect variation such as African American English [10], dialectal variation in modal constructions [11], and individual celebrity identity enactment through linguistic imperfection features [12,13]. Extensive research into vocabulary change has also been undertaken using Twitter corpora in Hong Kong and the Philippines [14], and the propagation of urban contact dialects such as Multicultural London English [15]. At a geopolitical level digital media reproduce ethnic and linguistic borders [16].

On the other hand, several studies have examined the relationship between Artificial Intelligence and language learning, particularly through bibliometric approaches that map instructional and pedagogical trends [17,18]. However, these approaches rarely engage critically with the sociolinguistic dimensions highlighted in language ideology research such as linguistic bias, identity representation, and ideological discourse reproduced by AI systems.

Regarding methodology, bibliometrics studies on AI are manifested in several fields, including operations management [19], e-commerce [20], renewable energy [21], and tourism [22]. In the area of linguistics, meanwhile, bibliometrics are more commonly applied to second language teaching and applied linguistics [23] and more broadly understood as the interconnection of Artificial Intelligence studies and linguistics [24]. However, the intersection of AI and sociolinguistics is still insufficiently researched, with few actual studies to account for. Furthermore, it is challenging to analyze this domain owing to the need for double-edged consideration of both scientific and ideological processes and phenomena, which necessitates a conceptual and methodological merger of quantitative and qualitative analytics.

This work attempts to address these gaps by the use of the term keyword-based content analysis with the relativity of the broad narrative content analysis offered by Braun and Clarke [25] and Klarin [26]. The base of keyword-based content analysis will be represented by the key unit of text that has discursive information about ideas presented in a scientific text. The interpretation of keyword co-occurrence focuses primarily on the representation of the co-word as a sign of the connection between the ideas and directions of research, the evolution of learning. This is a branch of framework-based approaches [27,28] that consider research as a complex structure to be read from the keyword level. Thus, the co-word analysis allows for interpreting the themes of the discourse through the interaction of keywords while using bibliometric data.

An initial literature search with the terms "artificial intelligence," "AI," "sociolinguistics," "bibliometric," and/or "scientometric" in Scopus and Web of Science revealed a lack of individual bibliometric studies on the intersection between AI and sociolinguistics. Literature currently available either concentrates only on A as a technique in AI, simply on sociolinguistic theory, or applies non-bibliometric methods. This demonstrates that we do not yet have a complete mapping of the combination of both domains.

Furthermore, an analysis of recent bibliometric studies in similar fields shows that most of them use only one database (Scopus or Web of Science) but not both. All the current analyses do not adopt a "double-search" policy

merging Scopus and WoS simultaneously in order to get the most exhaustive coverage. Accordingly, the joint use of these two well-known international databases in our study is a more abundant and comprehensive literature base compared with previous studies.

In addition to the absence of systematic mapping, previous bibliometrics have tended to be descriptive-quantitative and rarely accompanied by qualitative layers capable of reading the ideological meanings implied in the text [29,30]. To fill this gap, this study offers nine original contributions: First, it is one of the earliest bibliometric studies that has focused on the intersection between AI and sociolinguistics. Second, it integrates two well-recognized international databases (i.e., Scopus and Web of Science) for a strong search strategy to achieve complete literature coverage and high reliability. Third, it combines quantitative bibliometric methods with qualitative keyword-based content analysis to help identify scientific trends and collaborative networks as well as thematic dynamics and the language ideology embedded in the field. Fourth, the study demonstrates an undescribed epistemic shift from earlier technical research towards more ideological analyses of the question involving linguistic bias, representation, and identity. Fifth, revealing cross-cluster thematic connections allows us to see how technical, sociocultural, and ideological areas are woven together to form a repeated structurally interdisciplinary network of knowledge. Sixth, the finding of a territorial recentring of the epistemic Noeth and North American perspectives, complemented by contributions from regions such as Scandinavia and Southeast Asia. Several ideology-bearing terms frequently appear, including "identity," "bias," "representation," and "language ideology," suggesting the presence of linguistic power dynamics and potential bias within AI research discourse. Eight, the study traces a three-stage developmental trajectory in AI sociolinguistics, moving from initial technical exploration to a methodological phase and ultimately toward a more ideologically informed stage. Ninth, it positions computational sociolinguistic theory, an emerging field that has not been extensively documented in earlier literature.

Based on these objectives, this study is designed to answer the following four research questions:

- RQ1: What are the trends in the growth of publications on AI and sociolinguistics?
- RQ2: Who are the key actors—authors, institutions, and countries—and what are their collaboration patterns?
- RQ3: What are the main keywords that emerge, and how do their connections form the thematic structure of this field?
- RQ4: What research gaps have been identified, and where should future research be directed?

By foregrounding the ideological and cultural urgency of understanding AI-mediated language practices, this work does more than fill a gap in fragmented scholarship; it starts to generate a new dialogue about how AI and sociolinguistics are co-constituting one another in a new emergent landscape of knowledge production.

2. Materials and Methods

2.1. Research Approaches

This paper calls for a mixed-methods, exploratory design which adopts bibliometric analysis and keyword-driven qualitative interpretation to investigate interdisciplinary research at the intersection of Artificial Intelligence (AI) with Sociolinguistics in general. This choice reflects an effort to obtain a broad understanding of the knowledge structure, field maturation and ideological orientation in interdisciplinary AI–language research.

Bibliometric analysis was used to map publication growth, prolific authors, journals, countries, collaboration patterns, and keyword structures, providing an overview of the intellectual structure and evolution of the field [28,30]. The utility of bibliometric visualization for detecting thematic evolution and international collaboration dynamics in different scientific fields has been well established [31,32], whereas the incorporation of network analysis allows the identification of relational and dominance structures between communities [33].

To enrich this structural mapping, the paper used a keyword-oriented content analysis and limited it to author keywords extracted in the bibliometric phase. Descriptors were taken to be symbolic-epistemic devices that reflect conceptual preferences and ideological allegiances in scientific discourse [26,34]. Based on principles of co-word analysis, this method supplements quantitative bibliometric results with interpretative insights by systematically investigating main and co-occurrent keywords, which allows researchers to grasp the thematic and ideological dimensions of AI and Sociolinguistics studies in depth [27].

In addition to a bibliometric mapping, the current study applies keyword-based thematic interpretation to examine how conceptual patterns develop in the selected publications. This approach, however, does not correspond

to the full qualitative thematic analysis as outlined by Braun and Clarke (2006) [25] involving systematic coding and a close reading on a holistic level of full texts but rather works with interpretations of thematic signs taken from keywords.

More exactly, the relationships between author's keywords and co-word patterns as detected by a bibliometric tool are analyzed. Such keyword networks reveal repeated conceptual bundles and the changes in research interests over time, particularly on topics such as Computational Sociolinguistics, Natural Language Processing (NLP), ChatGPT, language, and machine learning. By focusing on relationships between keywords, the study seeks to identify wider thematic orientations across the literature, although also remaining in harmony with a bibliometric analytical approach. The focus thus is on locating conceptual trends at a broad, macro level rather than cloistering around in-depth qualitative interpretation of full-text data.

This method offers a more macroscopic view and is less sensitive to nuances of context from individual studies compared to traditional qualitative thematic analysis. Nonetheless, the keyword-based thematic interpretation offers a reasonable approach to review interdisciplinary growth and recognize emerging concept trends in this discipline.

2.2. Data Sources

The bibliographic information was obtained from Scopus and Web of Science (WoS) databases due to their trusted, standardized metadata and high multidisciplinarity coverage [35,36]. The information was limited to bibliographic metadata, such as titles, abstracts, author keywords, publication sources, author affiliations and citation metadata. These are valid for use in bibliometric mapping and content analysis based on keywords.

2.3. Article Identification

The search strategy was performed by accessing Scopus and Web of Science (WoS) databases with institutional accounts. Literature searches were conducted in article titles, abstracts and keywords to obtain comprehensive coverage of relevant studies. The search string was the Boolean search: ("artificial intelligence") AND ("sociolinguistics" OR "language ideology" OR "sociolinguistic"), which targeted at retrieving articles that were directly concerned with the interconnection between Artificial Intelligence and sociolinguistic research. Eligibility criteria for the documents included were: (1) academic publications, such as journal papers and conference papers; (2) materials published in English; and (3) studies detailing a focus on conversational agents and visual storytelling. Exclusion criteria involved review articles, books, book chapters and studies not written in English, as well as screening for references that did not fit within the thematic scope of the study. According to these criteria, 98 records were found at the first step of search (64 in Scopus and 34 in WoS) (**Figure 1**). The distribution of identified records by database is summarized in **Table 1**.

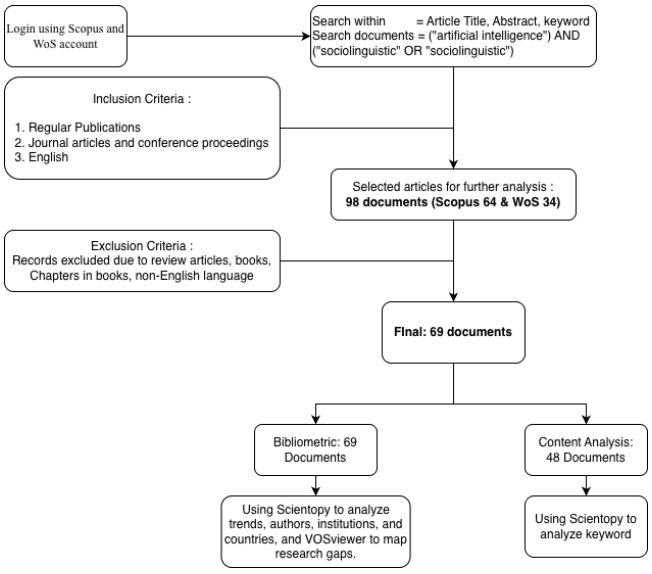

Figure 1. PRISMA-styled flow chart of the process for selection of articles at the level of abstract screening.

Table 1. Article Identification Stage.

Data Source	Records Identified
Scopus	64
Web of Science (WoS)	34
Total records identified	98

2.4. Screening and Deduplication

All records identified from Scopus and Web of Science (WoS) were merged and subjected to a systematic screening and deduplication process using ScientoPy. Duplicate identification was performed by an automated process that compared various menu citation fields (such as document titles, author names, published years, source titles and DOIs) to ensure that there were no duplicate records for a given publication. As shown in **Table 2**, a total of 98 records were screened 64 records from Scopus and 34 records from Web of Science (WoS). During deduplication, duplicate records were found and removed from Scopus dataset but no duplicate records were detected in WoS. A total of 35 from Scopus and 34 records from WoS were included as unique references (after removal of duplicates) [37,38]. In summary, the deduplication process yielded a final corpus of 69 unique documents, which were then analyzed bibliometrically. This screening phase was necessary for avoiding redundancy derived from cross-database indexing, circumventing the inflation of bibliometric indicators and also to guarantee the quality and reliability of further bibliometric mapping and content analysis based on keywords [39,40].

Table 2. Screening and Deduplication Results.

Source	Records before Screening	Duplicates Removed	Records after Deduplication
Scopus	64	29	35
WoS	34	0	34
Total	98	16	69

2.5. Inclusion for Bibliometric Analysis

Following the screening and deduplication process, all 69 unique documents were fully included in the bibliometric analysis. As shown in **Figure 2**, the findings of deduplication indicate that duplicate records were only found in the Scopus dataset and there was no record repeated in WoS [41,42]. Although some duplicate entries were initially available in Scopus, all records left after data cleaning satisfied the pre-determined inclusion criteria [43,44].

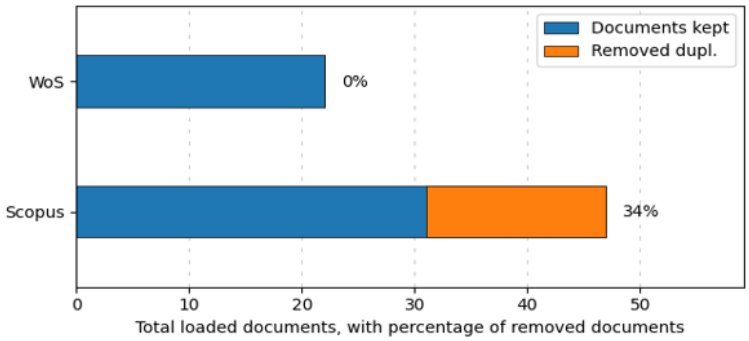

Figure 2. Deduplication results across Scopus and Web of Science datasets.

Consequently, we did not exclude any documents during the current step of screening as it also aimed at the elimination of duplicate records rather than relevance-based filtering. The resulting dataset of 69 unique documents therefore represents the complete bibliometric population, providing a comprehensive basis for mapping publication trends, authorship patterns, journal distributions, and keyword structures at the intersection of Artificial Intelligence and Sociolinguistics. All bibliometric analyses were subsequently conducted using ScientoPy [45,46].

2.6. Inclusion for Keyword-Based Content Analysis

Based on the bibliometric results specifically the dominant and recurrent keywords identified using ScientoPy—purposive subset of 48 full-text articles was selected for keyword-based content analysis. This choice did not entail exclusion by irrelevance, instead, it was a methodologically justified qualitative subsampling in order to allow deeper thematic and ideological interpretation of the most relevant concepts extracted from the bibliometric corpus.

2.7. Data Analysis Techniques

Bibliometric analysis was performed by ScientoPy to return descriptive statistics, including annual publication outputs, top authors and journals, country rankings, and keyword occurrence. The review was conducted descriptively to describe the temporal, spatial, and conceptual structure of the research field. The qualitative phase was carried out at the level of author keywords (as opposed to full-text documents) by conducting a keyword content analysis. Dominant keywords extracted from the bibliometric stage were manually coded and grouped to broader sociolinguistic themes in Microsoft Excel [47,48].

The interpretive grouping followed general principles of thematic organization, as outlined in Section 2.1, where thematic interpretation is conceptualized as a keyword-based analytical process rather than a full qualitative thematic analysis in the sense of Braun and Clarke (2006) [25]. Keywords were inductively clustered into groups based on their conceptual similarities and co-occurrences as observed during bibliometric analysis, thus enabling the capture of broader sociolinguistic issues while paralleling the analytical approach of bibliometry [49].

2.8. Methodological Limitation

The results should be interpreted considering a number of methodological limitations. The database contains only English-written papers, which could lead to language bias as well as when research is done in other languages not being presented. Despite the use of Scopus and Web of Science, which have standardized, high-quality metadata, using these databases could limit coverage to studies indexed elsewhere. Thematic interpretation with a keyword-based search, as opposed to full-text analysis, was conducted above the author keywords. This allows systematic mapping at the macro level but could lose the contextual details that are inherent in specific studies. The thematic grouping was done by one researcher who may have made the categories somewhat arbitrary though he made an effort to keep them consistent. Last but not least, bibliometric visualisation is accompanied by parameter choices (for instance minimum occurrences threshold) that structure the visibility of keyword clusters and should be seen as analytical interventions instead of as neutral representations of the research landscape.

2.8.1. Bibliometric Analysis

The bilingual analysis as presented in the previous study, was integrated with preliminary bibliometric data analysis using two main tools: ScientoPy and VOSviewer. Temporal trends, dominant themes and metadata filtering from the Scopus database were traced by ScientoPy [50]. This option was especially appreciated since it can return some key bibliometric expressions such as: (i) the major keywords being used, (ii) productive authors, (iii) leading countries and affiliations, and (iv) core journals within the field. It was instrumental in scientometric analyses to facilitate systematic mapping of research trends and thematic structures [51,52]. In this sense, van Eck and Waltman [53] underscore the capacity of ScientoPy to work with large sets of bibliographic data for the development of reproducible and structured trend analyses. VOSviewer [54,55] was used for visualization of bibliographic coupling, co-authorship networks and keyword co-occurrence to improve the interpretability of our bibliometric data. As an effective bibliometric tool, VOSviewer can be utilized to understand the structure of the scientific network, detect clusters in topics and visualize knowledge gaps and new trends [26]. Such capabilities helped shape evidence-based research agendas and facilitate strategic research directions that are academically and societally impactful [28].

The combined use of these bibliometric software tools has become a common practice and is popular [37] in safety knowledge literature is able to effectively exploit the duo of ScientoPy and VOSviewer. Further, the choice of Scopus and Web of Science as data sources is supported by Gazali and Saad [52], who claim that pooling these two databases provides a wider and more varied coverage of literature in cross-disciplinary bibliometric research. With this methodology the data was treated descriptively and visually to understand the dynamics of the study

development in terms of temporal, spatial and conceptual axes. The results of this stage thus form the basis for the next analytical procedure, i.e., keyword-based content analysis, which is focused on the meaning and thematic changes in the sense of co-words and dominant topics based on the keywords obtained.

2.8.2. Keyword-Based Content Analysis

Content analysis employed in this research was an experiment-oriented keyword-based content analysis method, which aimed to probe the conceptual meaning and ideological representation of the keywords that are listed in the literature (i.e., no reading of the article's full text). This choice is inspired by the assumption that keywords are interpreted as discourse nodes that model the epistemic structure (intent) and evolution of research in a given domain.

The investigation in this research was carried out with the help of two main techniques. We utilized ScientoPy to generate and search the top used keywords for temporal trends. The second stage of the keyword results from ScientoPy was to analyze these terms qualitatively in Microsoft Excel by coding and grouping them into sociolinguistic themes, namely identity, language ideology, linguistic representation and digital hegemony. This process was conducted through manual work alone, using the contextual information that this domain of research focuses on.

The keyword-based content analysis approach in this study is borrowed from the modern co-word analysis described by Braun and Clarke [25] and Klarin [26], analyzes how word co-occurrence patterns in a corpus can be used to systematically identify, cluster and interpret key concepts in terms of objective empirical rules, so that researchers can establish empirical thematic conceptual structures. This approach is justified by Klarin [26], who recommends combining keyword analysis and experts' conceptual analysis to increase the strength of a thematic analysis. Similar work was done by Olmeda-Gómez et al. [56] which maps the landscape of knowledge in a field through keyword cooccurrence and semantic clustering within a network structure representing discourse.

Related to this approach, this paper aims at interpreting keywords as discursive units reflecting meaning-making process, ideological bias and epistemic orientations in scientific literature of Artificial Intelligence (AI) and Sociolinguistics.

3. Results

3.1. RQ1: What Are the Trends in the Growth of Publications on AI and Sociolinguistics?

Figure 3 shows the temporal distribution of publications related to the themes of Artificial Intelligence and Sociolinguistics, based on data taken from two main databases, Scopus and Web of Science (WoS), during the period 1990 to 2024. The results of this bibliometric analysis suggest that interest in it from a scientific standpoint has started to gain momentum and concentrate after 2010 with respect to the published documents, particularly during the last decade. Publications retrieved from Scopus (blue circles) exhibit a variable but stable trend since 2010, with minor peaks in 2012 (3 papers), 2014 (4 papers) and 2023 (8 papers). This indicates that researchers from various disciplines have begun to focus their studies on issues at the intersection of AI and sociolinguistics, in line with the development of natural language processing technology and machine-based social data analysis.

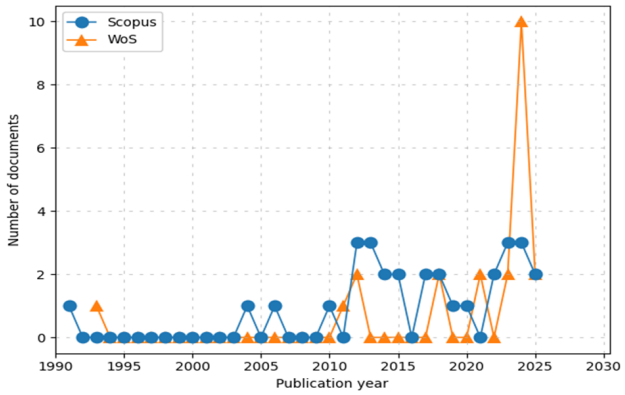

Figure 3. Trends in Artificial Intelligence and Sociolinguistics Publications in the Scopus and Web of Science Databases.

Meanwhile, publications from WoS (marked with orange triangles) show a relatively slow trend until 2021, but experience a sharp increase in 2024, reaching 10 reaching 10 WoS-indexed publications in one year. **Figure 3** has also been resized to maintain proportional scaling and improve readability. This peak signifies a surge in global academic attention to the integration of artificial intelligence and sociolinguistic studies, which may correspond with the global attention toward generative AI development and its linguistic applications rather than directly resulting from it.

Scopus' dominance throughout most of the period indicates that many interdisciplinary publications in this field are still predominantly found in conference proceedings and technical journals, which are more extensively indexed in Scopus. The significant rise of WoS articles in recent years, nevertheless, might indicate that literature on this theme is starting to be incorporated in journals with a more rigorous academic selectivity and accumulated scientific impact.

The upward slope in both databases suggests that Computational Sociolinguistics is becoming an established interdisciplinary field of inquiry within contemporary linguistics. This bibliometric trend indicates the growing interconnectedness of AI and sociolinguistics as two strategic and interdisciplinary domains. The rise in publications not only reflects a growth of academic interest, but also indicates great potential for theoretical inquiry and ethical reflection in this area, especially regarding the development of technology and social consciousness concerning the impact of language on AI.

3.2. RQ2: Who Are the Key Actors—Authors, Institutions, and Countries—And What Are Their Collaboration Patterns?

To answer RQ 2, see **Table 3** and **Figure 4**. **Table 3** shows who the leading authors are (Strzalkowski T., Shaikh S., etc.), how active their contributions are, and their bibliometric metrics.

Table 3. List of Most Productive and Influential Authors in Artificial Intelligence and Sociolinguistics Studies.

Pos.	Author	Total	AGR	ADY	PDLY	h-Index
1	Strzalkowski T.	5	0.0	0.2	40.0	4
2	Shaikh S.	4	0.0	0.1	25.0	4
3	Boz U.	3	0.0	0.0	0.0	3
4	Broadwell G.A.	3	0.0	0.0	0.0	3
5	Liu T.	3	0.0	0.0	0.0	3
6	Stromer-Galley J.	3	0.0	0.0	0.0	3
7	Taylor S.	3	0.0	0.0	0.0	3
8	Ravishankar V.	2	0.0	0.0	0.0	2
9	Ren X.	2	0.0	0.0	0.0	2
10	Briggs G.	1	0.0	0.0	0.0	1

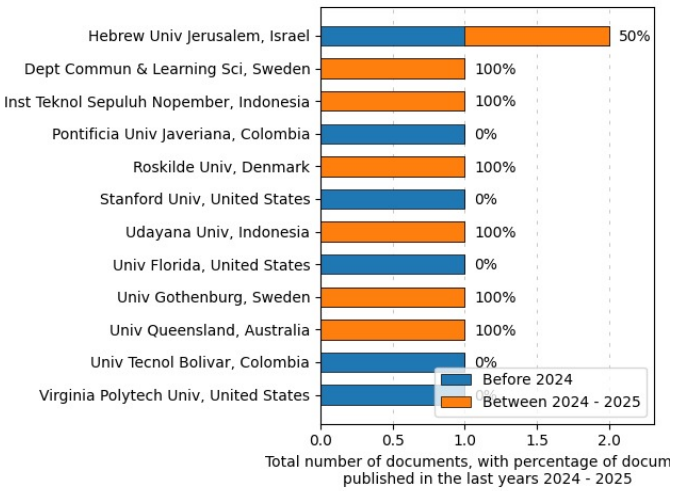

Figure 4. Institutional and National Distribution of Publications on Artificial Intelligence and Sociolinguistics and Percentage of Recent Documents (before 2024 vs. 2024–2025).

Table 3 presents the ten most productive and influential authors in Artificial Intelligence and Sociolinguistics research, showing that Strzalkowski T. ranks first with five publications (h-index = 4), followed by Shaikh S. with four publications (h-index = 4), both of whom have played a pivotal role in developing computational sociolinguistics and AI-mediated discourse analysis. Meanwhile, a second group comprising Boz U., Broadwell G.A., Liu T., Stromer-Galley J., and Taylor S., each producing three documents (h-index = 3), contributes to studies on language variation, human–AI communication, and digital discourse, whereas Ravishankar V. and Ren X. authored two papers each (h-index = 2), and Briggs G. produced one publication (h-index = 1). Although all authors record zero Annual Growth Rate (AGR) and low Average Documents per Year (ADY), the Percentage of Documents in the Last 10 Years (PDLY)—40% for Strzalkowski and 25% for Shaikh—indicates that recent research activity is increasingly concentrated among leading contributors. In general, the results presented also confirm in this case how Strzalkowski T. and Shaikh S. dominate over the corpus while other scholars provide sporadic and different contributions, revealing an emerging but not yet stabilized trend going across AI literature and sociolinguistics itself as a growing interdisciplinary field, not yet measure that is developed through collective and focused interest, rather than individual dominance.

Figure 4 illustrates the distribution of institutional contributions across countries, distinguishing publications produced before 2024 and between 2024 and 2025. The data reveal growing global diversification in AI–Sociolinguistics research. Institutions such as the Department of Communication and Learning Sciences (Sweden), Institut Teknologi Sepuluh Nopember (Indonesia), Roskilde University (Denmark), Udayana University (Indonesia), University of Gothenburg (Sweden), and University of Queensland (Australia) show 100% of publications between 2024 and 2025, indicating an emerging and active research engagement. The Hebrew University of Jerusalem (Israel) maintains continuity with 50% of its outputs in the recent period, reflecting sustained scholarly involvement.

Conversely, earlier contributions are dominated by Pontificia Universidad Javeriana (Colombia), Stanford University, University of Florida, Universidad Tecnológica de Bolívar, and Virginia Polytechnic Institute (United States), all showing no new publications in 2024–2025. This trend is indicative of a movement away from North American and Latin American centrality, towards the emergence of research activity in Scandinavia, Southeast Asia and Australia as part of a more general transnational expansion phenomenon for the AI–sociolinguistics epistemic network.

3.3. RQ3: What Are the Main Keywords That Emerge, and How Do Their Connections Form the Thematic Structure of This Field?

Figure 5 shows the distribution of the five main keywords based on year of publication. The keyword Computational Sociolinguistics has the longest time span, starting in the early 1990s and continuing through the mid-2020s, but only one document or fewer is recorded each year. Natural Language Processing stands out in 2021 with the highest number of documents among all keywords in a given year. The keyword ChatGPT appears in the last two years on record, namely 2024 and 2025. Machine Learning is identified at several scattered points in time, namely around 2013, 2014, and 2024. Meanwhile, Language only appears once, namely in 2019. This graph shows that each keyword has a different pattern of appearance, both in terms of time and the number of related documents.

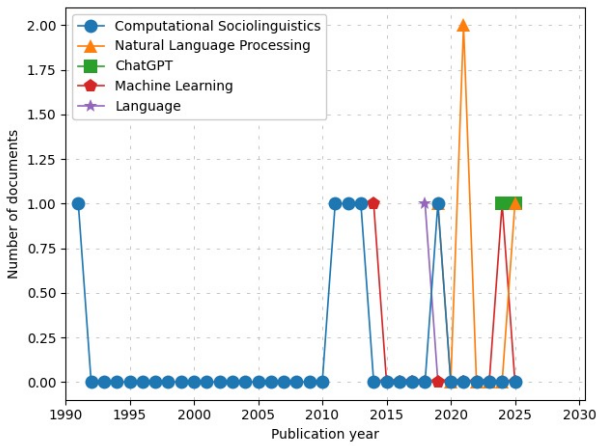

Figure 5. Visualization of Keywords in Artificial Intelligence and Sociolinguistics.

In addition to frequency and year of presence, the keyword co-occurrence analysis also clarifies the differentiated thematic clusters in configuring the knowledge structure. Cluster 1 (red) includes terms like Computational Sociolinguistics, Language Ideology and Bias, showing the epistemological and ideological aspects of language in AI. Cluster 2 (green): comprising Natural Language Processing, Machine Learning and ChatGPT, which is referred to as the technological and algorithmic core of AI-based linguistic research. Cluster 3 (blue) features terms such as Discourse Analysis, Identity Construction, or Multilingualism, which are indicative of the sociocultural trend in current research. Thematic Markers of intersection between these clusters imply a dialectical pattern: computational methods are now being applied to the study of sociolinguistic phenomena, and ideological and ethical debates follow from these computational applications. This is an example of how the language study of AI itself is changing from a strictly technical analysis to more analytic and socially anchored questioning.

The field is seen in clear temporal terms to move from cataloguing intellectual and technical issues (2010–2018) to ideological and ethical ones (2019–2025). The introduction of ChatGPT and similar generative models has hastened interest in language ideology, bias, and representativeness indicating a wider epistemological shift within computational sociolinguistics.

3.3.1. Computational Sociolinguistics

Computational Sociolinguistics appears to have become more significant in corpus as a keyword, operating in an explicit and implicit manner either as a methodological platform or an epistemological mediator between the computational model and the social analysis of language. In the context of keyword-based content analysis, it mirrors academic teeth gnashing about how social relations are reified through statistical model engineering linguistics features. This framework has been utilized by several works that incorporate sociolinguistic variables into AI dialogue modeling. For example, interactional attributes such as authoritativeness, heteroglossia, and power-relevant phenomena—such as topkill dominance and disagreement—have been proposed to account for the complexity of multi-party communication [57]. Furthermore, research that employed both linguistic and sociocultural cues investigated gender specific properties in social media discourse [58]. One step further, Grieve et al. [59] suggest that (Large) Language Models (LLM) are artifacts of computational sociolinguistics, since social language varieties are encoded in LLMs and thus linguistic justice principles should be embedded already at the model design level. Similarly, Abitbol et al. [60] and Tarrade et al. [61] use French-language Twitter data to show the correlations between linguistic patterns and socio-economic status of users and 24 illustrate how lexical novelties are spread through digital networks using big data. Departing from, but not dismissing entirely, critical readings of bias and dominance, Nissan [62] points to the performing capacities of this paradigm through the ONOMATURGE system which uses AI to save national languages by creating new lexical items for marginalized linguistic communities. They also show that Computational Sociolinguistics appears not only as a technical–methodological field but as an epistemological commitment to the inclusion of sociolinguistic relativism in AI modelling, by establishing an attending discussion on linguistic justice, language ideology and equitable representation of language for the digital era.

3.3.2. Natural Language Processing (NLP)

The term NLP (natural language processing) appears frequently in examined corpora and implies a good-minded concept integration in the AI interdiscipline with sociolinguistics. Through the keyword content analysis, NLP is approached not only as a device for technically parsing language but also an ideological instrument that reflects and replicates contested sociolinguistic systems. Multiple papers discuss how NLP systems tend to introduce bias towards non-standard language idiosyncrasies [63,64], where processing of dialectal and casual orthography especially on social media poses a challenge for most systems so that they favour standard linguistic forms. Similarly, Astuti and Sari [65] emphasize that NLP-based speech recognition often misunderstands African American Vernacular English (AAVE) and supports digital exclusion through the linguistic "othering" process.

In the Indonesian context, Guo et al. [66] emphasize NLP approaches' incapability to address widespread code-mixing on regional social media; it demonstrates how mainstream systems cannot tailor themselves to sociolinguistic variation in the Global South. These local findings resonate with wider critiques that many NLP systems represent majority linguistic ideologies while sidelining minority and underrepresented languages. For instance, Waliya [67] and Tran and Stell [68] posit that NLP methodologies encode normative assumptions about language use and spread discrimination patterns or ignore demographic properties of language production among minor-

ity users. Similarly, Tran and Stell [68] comment that those tools are frequently constructed under those strong assumptions on linguistic uniformity, ignoring pragmatic, contextual and cultural variation. In practice, NLP is often used to predict social traits such as demographics [4], gender [5], and personality [6] from textual data; in this mode of application, the computer is provided with samples of human language and makes predictions about human social properties. For example, Moreno-Sandoval et al. [13] employ NLP in the exploration of lexical polysemy in Colombian academic discourse at the university level, displaying two faces, inspired and reductive. In the clinical space, Guo et al. [66] underscore that while NLP-based large language models are promising as conversational agents in mental health care, they still have the potential to propagate semantic and ideological bias via predictive language modeling. In general, NLP is not a neutral linguistic mechanism: it encodes epistemic and ideological assumptions that regulate how linguistic norms and power relationships are mediated in digital environments.

3.3.3. ChatGPT

The keyword ChatGPT becomes heavily prominent within the analyzed corpus, specifically after 2023 when it is quickly accepted into interdisciplinary conversations about AI and Sociolinguistics. In the analysis of a keyworded corpus, ChatGPT is not merely seen as a technological object but rather as a discursive one that reifies sociolinguistic stratification. In current research it has been described as both a linguistic generator and a normative actor that constrains perceptions, values, and communicative authority in the digital world at large. Empirical work [69] shows that ChatGPT systematically projects biases from its training on Standard English, and tends to produce responses that stereotype or suppress other serviable varieties. Also, Yibokou et al. [70] find that its partial inclusion of Vietnamese and Mandarin dialectal variants consolidates the dominance of standard languages and threatens a linguistic 'recolonization' in educational settings. Beyond linguistic priming, it impacts discourse authorship and pedagogy: ChatGPT's linguistic framing biases credibility appraisal of crisis information [71], while English language learners' (ELL) stylistic patterns are affected by structural dominance, e.g., participial clauses [72]. Similarly, Xiao and Yu [69] and Sanei [73] caution that wholesale incorporation of ChatGPT into EFL and L2 educational contexts might sound the death knell for indigenous linguistic traditions along with culturally based pedagogies. Simultaneously, as conversations unfold, it seems that while ChatGPT is a limited medium for cultural nuance, when critically and contextually applied to pedagogy—it holds potential for flexible learning support. Overall, the discussion suggests that ChatGPT is not only a language generator but also an ideological agent that reshapes linguistic norms, reallocates communication authority, and makes dominant variants of speech more prestigious. However, recognizing these constraints also paves the way for creating culturally adaptive and inclusive AI technologies that support linguistic diversity as well as equitable representation across geopolitical contexts.

3.3.4. Language

As a result, we select "Language" as the keyword which is of high frequency and more multidimensional to connect Artificial Intelligence (AI) with sociolinguistic discourse. In a keyword-based analysis of content, domain "language" covers the ideological spectrum from representation and social categorization to linguistic exclusion in technological contexts. Researchers have discussed the ways in which language does not serve as a barrier, but rather an instrument of power and social signification, in digital contexts. Laitinen et al. [74] argue that on the Internet, language use replicates social inequalities and as Laitinen and Lundberg [75] stress the normative character of Internet orthography in which language is a field of power and agency enacting discursive chronotopes. Similarly, Goel et al. [76] and Gonzales [77] position language as a social marker that is traded over digital networks with social mobility and network connectivity affecting rates of linguistic adoption and change.

Syntactic variation and form selection are thus seen as markers of social identity construction [78,79]. While Dijkstra et al. [80] and Alshaabi et al. [81] portray digital communication as a site of contestation between linguistic survival and extinction**,** especially for endangered and minority languages. But language also possesses affective and expressive qualities; Krishna et al. [82] demonstrate how it functions as a medium of emotional reference while also providing grounds for insult within online contexts. The centrality of English has frequently been associated with the increasing homogenization of language practices globally, determining patterns of persuasion and the organization of communication across space. In educational and interactive settings, language is not merely a tool of communication; it is also a strategic resource for learning, translation, and social encounter. Taken together, these perspectives suggest that language operates as a political and ideological space, where algorithmic systems in AI

and NLP shape visibility, legitimacy, and inclusion within global digital communication.

3.3.5. Machine Learning

The keyword "Machine Learning (ML)" occupies a key position in this analysis, particularly gaining prominence in the post-2020 corpus. In keyword analysis of content, ML encompasses not just the computational methods used to manipulate textual data but also the procedural practices that manage how linguistic identities and social categories are classified. A number of studies emphasize the importance of interdisciplinary relationships between language and technology in helping to develop computational approaches for processing linguistic data [83]. Following up on this research, work in the study of social media environments (notably Twitter) has shown that machine learning can use lexical and interaction features to detect gender profiles [84] as well as anonymous accounts [22]. Taken as a whole, these studies demonstrate how the models can be used to transform patterns of linguistic behaviour into numerical measurements of social identity.

Unfortunately, this precision often comes at the price of contextual knowledge. Simaki et al. [85] point out that ML models lack the "human context", while, Tarrade et al. [61] critically note that NLP and ML paradigms consider language as neutral statistical data and disregard its sociocultural and ideological implications. Technically, Devi and Sharma [86] show that deep learning models can distinguish Arabic dialects within multilingual corpora with high accuracy (and are sensitive to phonetic and linguistic diversity) while Barakat et al. [87] and Glazkova et al. [88] find that document level lexical cues increase predictive accuracy in age-based text classification exemplifying how technical precision may blind us to social nuance.

Taken together, in digital sociolinguistics "machine learning" operates as a methodological tool and an ideological technology of social classification that may unwittingly rigidify linguistic diversity, user agency, and non-quantifiable forms of the social. These models not only predict language but also naturalize identities that are formed by statistical prejudice. For this, we have to follow (ML) applications in linguistic questions with a critical consideration between precision and the social meaning of such a kind of algorithmic representation as Glazkova et al. [88] warn for.

3.4. RQ4: What Research Gaps Have Been Identified, and Where Should Future Research Be Directed?

The bibliometric profiling offered in the previous section offers a macro-view on the topography of Artificial Intelligence (AI) and sociolinguistic enquiry. However, in order to go beyond these trends of publication, a qualitative approach is necessary that enables us to delve into the discursive constructions that make these trends possible, address what meaning is being produced and what is being contested, and what the ideology and power relations that are intrinsic to these scientific forms of knowledge. For this purpose, the paper is going to rely on a content analysis of digital media content, utilizing keywords as entry points into the discursive formations that might constitute the intersections of language, power and technology drawing upon keyword work.

This is how we see keywords at links, not as thematic indices, but the conceptual nodes that represent particular epistemological interests (and often strategic battlegrounds of negotiation and meaning). For this reason, the five most dominant keywords in the bibliometric network namely Computational Sociolinguistics, Natural Language Processing (NLP), Machine Learning (ML), ChatGPT, and Language are critically examined to reveal how technology and language are constructed, disciplined, and positioned within contemporary academic discourse. Through this analysis, the study aims to demonstrate that linguistic dynamics within the domain of AI are far from neutral; instead, they reflect complex social, cultural, and ideological configurations.

Figure 6 is a summarizing diagram of the key research gaps discovered in AI and Sociolinguistics for which directions for future research are proposed. The figure consolidates both the bibliometric and content analysis results, identifying six main axes where more attention should be devoted by scholars, including, for example, the ideological dimension in NLP studies, or the presence of minority languages in our core set. It indicates how methodological innovation, rather than just descriptive methods should underpin future empirical work, harnessing critical, ethical and interpretive frameworks. In particular, it provides a conceptual link between the empirical mapping of the field and the normative agenda to promote socially responsible and inclusive AI research.

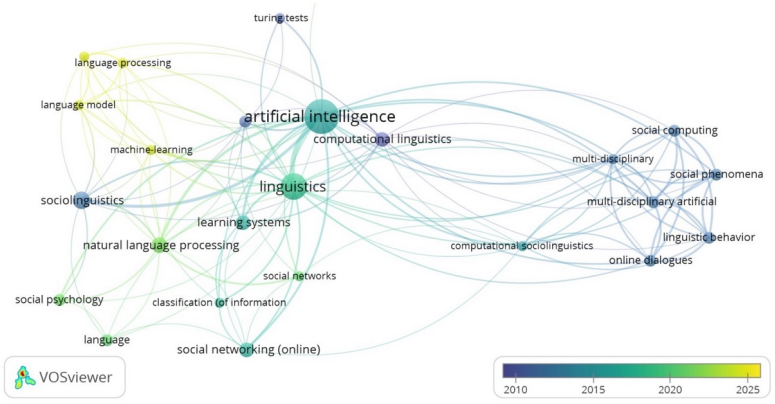

Figure 6. Research Gaps and Recommended Future Directions.

Table 4 presents a detailed correspondence between six identified research gaps and their associated recommendations for future research directions. The first three gaps overemphasize on representational bias, lack of focus on minority languages, and insufficient sociolinguistic theorization reflect epistemological limitations in current AI–linguistic studies. Gaps between technical and humanities work, in lack of attention to user agency and quantitative positivism, as well as fragmentation represent methodological challenges for the discipline. The table offers practical recommendations, including the inclusion of critical sociolinguistic approaches, promotion of interdisciplinary collaborative effort and application of combined qualitative–quantitative fieldwork methods. Overall, **Table 4** serves as an integrated roadmap for future research to grow into a more inclusive, reflexive and theoretically coherent field of AI-mediated language practices.

Table 4. The research gaps and recommended future directions.

No.	Research Gaps	Recommended Future Research Directions
1.	An overemphasis on bias and representation issues in NLP without deeper ideological exploration	In-depth investigations into how AI internalizes and disseminates dominant language ideologies through algorithmic structures
2.	A lack of research on minority languages, local languages, and multilingual communities	Prioritize the representation of marginalized languages in AI development, and examine its implications for language preservation and homogenization
3.	Absence of critical sociolinguistic perspectives in analysis	Integrate critical sociolinguistic theories to examine power relations embedded in AI design and output
4.	Fragmentation between technical (AI) and humanities (sociolinguistic) approaches	Foster interdisciplinary collaboration among NLP engineers, sociolinguists, and digital humanities scholars
5.	Limited focus on user agency in AI-mediated interactions	Conduct studies on how users adapt to, negotiate, or resist linguistic forms embedded in AI systems
6.	Predominance of descriptive-quantitative methodologies	Combine bibliometric analysis with content analysis, discourse analysis, and qualitative approaches for more reflective and context-sensitive insights

4. Discussion

This research highlights the complex interrelation between artificial intelligence (AI) and sociolinguistics from methodological, epistemological and ideological perspectives. The paper looks at four main research questions: RQ1 (publication trends); RQ2 (top actors and collaborations); RQ3 (ideological representations by keywords) and RQ4 (research gaps and future study directions). It pools Scopus and Web of Science (WoS), and involves quantitative as well as qualitative analysis. This technique is a new approach to enhance the comprehensiveness of data coverage, improve the accuracy in analysis and enrich the understanding of literature interpretation.

4.1. RQ1: What Are the Publishing Trends Concerning AI and Sociolinguistics from 2013 to 2024?

Otherwise, the bibliometric analysis using dual-database integration (Scopus and Web of Science) shows a growing interest from 2013 to 2024 for publications at the crossroad between Artificial Intelligence (AI) and Sociolinguistics with an upward trend forecast that this should increase since 2020. This increase reflects the increasing popularity of language-based AI applications like Natural Language Processing (NLP) and Large Language Models (LLMs). Generative AI, as observed by Hagos et al. [89], enables us to do a statistically massive inquisitive linguistic

task. These developments also illustrate that language technologies interact ever more actively with sociolinguistic practices, indicating that the AI-language interface has crucial ideological dimensions.

While the specific term "Computational Sociolinguistics" is still not extensively used, its concept becomes more widespread. Computational framings of linguistic influence and power, as seen in the studies of Blodgett et al. [9] and Habernal and Gurevych [90], can create empirical connections between language and data on one hand, and social structure on the other. In a similar vein, Strzalkowski et al. [91] show that the linguistic variation in digital platforms is actually indicative of underlying social dependencies, once again attesting to the fact that AI-based methods have become sociological tools for unearthing social meaning from language traces.

Geographically, most publications are still coming from Global North (United States, United Kingdom and China), but countries in Southeast Asia especially Indonesia and Malaysia, have been more active since 2020. Implications This trend demonstrates on the one hand the continuing asymmetry in (knowledge) production at global level and ambition on the other of epistemic diversity that is already beginning to assert itself with this study, which methodologically is also able to fill a gap by combining quantitative bibliometrics and qualitative keyword-based content analysis in two steps so as not only to record but bring into view—using both structural and ideological dimensions. The general direction of travel suggests that AI and Sociolinguistics are not only methodologically intersecting, but ideologically: the specter of computational tools is increasingly mediating how language, identity and social power are enacted in digital spaces.

4.2. RQ2: In What Manner Do the Primary Keywords Embody Theme Frameworks and Discursive Perspectives?

By examining keywords "Natural Language Processing (NLP)," "Machine Learning (ML)," "ChatGPT," and "Language", topical patterns along with stark ideological postures that AI may possess when it encounters the practice of language are also uncovered. These keywords are discursive signifiers which contain encoded value systems about language, power and representation in digital environments. These are not just technologies but as we will see come to represent competing epistemologies across computational sociolinguistics: from the modeling of linguistic patterns quantitatively to critically reflecting on language as a site of social inequality and digital marginalization.

Recent literature has raised growing concerns about the assumption that natural language processing (NLP) and machine learning are neutral tools, highlighting how they often reproduce central linguistic patterns in their training data. Research on generative AI in academic and scientific writing shows how linguistic monolingualism and standardization are often favored, posing issues of bias and epistemic homogeneity in scholarly communication [92, 93]. Past studies also show that algorithmic bias is the result of under-representation of non-standard language varieties and minority speech patterns, which further replicate already existing linguistic hierarchies in computational systems [94]. These trends also appear in analyses of standard language ideology in AI- generated texts where linguistic homogeneity can tacitly present itself as correctness or quality [95]. In this way then, AI is not only the processor of language, but in fact a participant in the building of linguistic authority by encoding patterns that resonate with existing ideological prejudices.

4.3. RQ3: Ideological Aspects of Keywords

The socio-political nature of language processing as opposed to its undeniable neutrality has become evident through the ideological dimensions present in AI-driven linguistic technologies. As noted by Blodgett et al. [9], these are merely 'technical issues with representation' in NLP, and not the only source of bias; rather, they also manifest in the reproduction of social hierarchies and ideological beliefs around language use. In the analysed corpus, NLP and ML are epistemic tools identifying the linguistic data, while maintaining contemporaneously normative régimes of "acceptable" language. This is particularly evident in the way that NLP models systemically misrepresent non-standard styles, dialects or code-mixed speech, and with it erase alternative linguistic forms beyond dominant norms.

More generally, Dunn and Edwards-Brown [96] demonstrate that language models like ChatGPT turn variation into ordered evaluations, in which proximity to Standard English becomes synonymous with credibility and accuracy. Such algorithmic assessments reproduce what Astuti and Sari [65] term "sociolinguistic encoding": the nesting of social meaning within computational representations of language. What results then is an unconscious

ideology that valorizes full language users and disempowers those communities, that speak with a linguistic repertoire beyond standard conventions.

From a more general point of view, by way of this analysis we can see how AI technologies both translate and regulate language through algorithmic classification. As Doval et al. [97] underline, computational normalization of linguistic data sorts and classifies users in geographic and phonetic patterns consolidating power disparities. Thus, AI-mediated language use has ideological effects that entangle issues of representation, access and epistemic justice, underlining the fact that digital linguistics is a domain through which social power and computational authority intersect.

4.4. RQ4: Research Deficiencies and Future

Despite progress in AI-based language tools, crucial gaps exist, in relation to inclusivity and linguistic diversity as well as socio-linguistic grounding. Data-based models often reduce complex social categories like class, race and region to statistical abstractions which in turn can serve to sustain existing hierarchies. Yet the growing digital language gap is further marginalizing under-resourced languages in standard NLP development. These trends mirror larger data access and research priorities imbalances that perpetuate the hegemony of Global North linguacentric perspectives.

In addition, very little literature considers sociolinguistic theory when implementing a computational model. As Curry et al. [98] note, much work in NLP and ML considers bias as a technical feature of a model rather than as a social artefact, failing to take account of how ideology will shape what is considered to be 'fair' or 'accurate'. Abitbol et al. [60] also lambast the sociolinguistics-analytic tradition in relation to computational modeling, for being non-reflexive and utilitarian. These lacunae point to the necessity for a more radical theoretical engagement that apprehends language as a symbolic order and a scene of ideological contestation.

However, in future work a holistic approach that combines computational accuracy with sociolinguistic intuition should be pursued. This requires creating AI models in support of minority and low-resource languages so that inclusivity is by design rather than an afterthought. More generally, a successful linkage of technically innovative with social accountability could help research on AI to grow from being strictly descriptive modeling theories toward transformative practice in which digital linguistics serves as not just an instrument of technological improvement but also as a force for linguistic justice and epistemic diversity in the world knowledge ecosystem.

5. Conclusions

This research maps knowledge in the intellectual and thematic space between Artificial Intelligence and Sociolinguistics using bibliometric analysis, keyword-based content analysis, and a combination of Scopus and Web of Science. This dual-database, dual-method approach constitutes a new methodological strategy that combines quantitative trend analysis with qualitative discourse interpretation in order to capture both structural and ideological change within the field. Findings indicate the increasing academic interest in this nascent interdisciplinary area by researchers, focusing on Computational Sociolinguistics, NLP, ChatGPT and language and machine learning. The two levels of investigation, bibliometric and content analysis, in combination point to how AI-facilitated research mirrors dynamic epistemic paradigm transformations away from technically modelled language towards questions about the nature of language as a sociocultural object enmeshed within digital systems.

This binocular way of analyzing does not simply chart the state of the art in the field, it reveals voids and emerging areas, particularly with regards to sociolinguistic implications of AI. It brings to the fore, the necessity for socially responsible and linguistically diverse AI systems that enable epistemic inclusivity and linguistic fairness in computational work. Bringing together bibliometric precision with sociolinguistic interpretation, the study offers a fuller understanding of how power, ideology and representation are implicated in AI-mediated communication.

In closing, this study pushes for a future-facing agenda that sees AI not only as a technology but as an increasingly sociolinguistic form of discourse and identity and inclusion in the digital age. It provides a foundation to outgrow and rethink an angle of deploying technology that can nurture fairer, multilingual, human-centered communication ecosystems in the future.

Author Contributions

Conceptualization, R. (Rugaiyah) and N.G.; methodology, R. (Rugaiyah) and A.I.; formal analysis, A.I.; writing—original draft preparation, R. (Rugaiyah); writing—review and editing, R. (Rugaiyah); supervision, R. (Roziah); funding acquisition, N.S. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Directorate of Research, Technology, and Community Service (DRTPM) under grant number 138/C3/DT.05.00/PL/2025.

Institutional Review Board Statement

Not applicable. This study did not involve human participants or animals and used publicly available bibliometric data from Scopus and Web of Science databases accessed via Universiti Utara Malaysia.

Informed Consent Statement

Not applicable.

Data Availability Statement

This article is made possible by the access provided by Universiti Utara Malaysia to the Scopus and Web of Science (WoS) databases, which were used to obtain the bibliometric data.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the Ministry of Higher Education, Science, and Technology of Indonesia (Kemdiktisaintek) for funding this research.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

During the preparation of this manuscript, AI-assisted tools such as QuillBot and DeepL were used solely for language editing, paraphrasing, and translation purposes. These tools did not contribute to the study design, data analysis, interpretation of results, or scientific conclusions. The authors take full responsibility for the intellectual content of this manuscript.

References

1. Lucy, L.; Bamman, D. *Gender and Representation Bias in GPT-3 Generated Stories*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 48–55. [CrossRef]
2. Sheng, E.; Chang, K.-W.; Natarajan, P.; et al. The Woman Worked as a Babysitter: On Biases in Language Generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, November 2019; pp. 3407–3412. [CrossRef]
3. Blasi, D.; Anastasopoulos, A.; Neubig, G. Systematic Inequalities in Language Technology Performance across the World's Languages. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, May 2022; pp. 5486–5505. [CrossRef]
4. Joshi, P.; Santy, S.; Budhiraja, A.; et al. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, online, July 2020; pp. 6282–6293. [CrossRef]
5. Kreutzer, J.; Caswell, I.; Wang, L.; et al. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 50–72. [CrossRef]

6. Ruder, S.; Peters, M.E.; Swayamdipta, S.; et al. Transfer Learning in Natural Language Processing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, June 2019; pp. 15–18. [CrossRef]
7. Hovy, D.; Prabhumoye, S. Five Sources of Bias in Natural Language Processing. *Lang. Linguist. Compass* **2021**, *15*, e12432. [CrossRef]
8. Helm, P.; Bella, G.; Koch, G.; et al. Diversity and Language Technology: How Language Modeling Bias Causes Epistemic Injustice. *Ethics Inf. Technol.* **2024**, *26*, 8. [CrossRef]
9. Blodgett, S.L.; Barocas, S.; Daumé, H.; et al. Language (Technology) is Power: A Critical Survey of 'Bias' in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020; pp. 5454–5476. [CrossRef]
10. Jones, T. African American English Intensifier Dennamug: Using Twitter to Investigate Syntactic Change in Low-Frequency Forms. *Front. Artif. Intell.* **2023**, *5*, 683104. [CrossRef]
11. Morin, C.; Desagulier, G.; Grieve, J.A.C.K. A Social Turn for Construction Grammar: Double Modals on British Twitter. *Eng. Lang. Linguist.* **2024**, *28*, 275–303. [CrossRef]
12. Puertas, E.; Moreno-Sandoval, L.G.; Redondo, J.; et al. Detection of Sociolinguistic Features in Digital Social Networks for the Detection of Communities. *Cogn. Comput.* **2021**, *13*, 518–537. [CrossRef]
13. Moreno-Sandoval, L.G.; Pomares-Quimbaya, A.; Alvarado-Valencia, J.A. Celebrity Profiling through Linguistic Analysis of Digital Social Networks. *Comput. Soc. Netw.* **2021**, *8*, 16. [CrossRef]
14. Gonzales, W.D.W. Broadening Horizons in the Diachronic and Sociolinguistic Study of Philippine English with the Twitter Corpus of Philippine Englishes (TCOPE). *Eng. World-Wide* **2023**, *44*, 403–434. [CrossRef]
15. Ilbury, C.; Grieve, J.; Hall, D. Using Social Media to Infer the Diffusion of an Urban Contact Dialect: A Case Study of Multicultural London English. *J. Socioling.* **2024**, *28*, 45–70. [CrossRef]
16. Demaj, U.; Vandenbroucke, M. Persistence of Ethnic and Linguistic Division During the COVID-19 Pandemic Outbreak in Kosovo. In *COVID-19 and a World of Ad Hoc Geographies*; Springer: Cham, Switzerland, 2022; pp. 2361–2379. [CrossRef]
17. Rahman, A.; Raj, A.; Tomy, P.; et al. A Comprehensive Bibliometric and Content Analysis of Artificial Intelligence in Language Learning: Tracing between the Years 2017 and 2023. *Artif. Intell. Rev.* **2024**, *57*, 107. [CrossRef]
18. Liu, C.-Y.; Yin, B. Affective Foundations in AI-Human Interactions: Insights from Evolutionary Continuity and Interspecies Communications. *Comput. Hum. Behav.* **2024**, *161*, 108406. [CrossRef]
19. Dhamija, P.; Bag, S. Role of Artificial Intelligence in Operations Environment: A Review and Bibliometric Analysis. *TQM J.* **2020**, *32*, 869–896. [CrossRef]
20. Bawack, R.R.E.; Wamba, S.F.; Carillo, K.D.A.; et al. Artificial Intelligence in E-Commerce: A Bibliometric Study and Literature Review. *Electron. Mark.* **2022**, *32*, 297–338. [CrossRef]
21. Zhang, L.; Ling, J.; Lin, M. Artificial Intelligence in Renewable Energy: A Comprehensive Bibliometric Analysis. *Energy Rep.* **2022**, *8*, 14072–14088. [CrossRef]
22. Knani, M.; Echchakoui, S.; Ladhari, R. Artificial Intelligence in Tourism and Hospitality: Bibliometric Analysis and Research Agenda. *Int. J. Hosp. Manag.* **2022**, *107*, 103317. [CrossRef]
23. Kartal, G.; Yeşilyurt, Y.E. A Bibliometric Analysis of Artificial Intelligence in L2 Teaching and Applied Linguistics between 1995 and 2022—Addendum. *ReCALL* **2024**, *37*, 441. [CrossRef]
24. Yaseen, M.G.; Alkattan, H.; Farhan, L. The Evolution of Computational Linguistics: A Bibliometric Analysis of Research Trends from 1966 to 2023. *Appl. Data Sci. Anal.* **2025**, 2025, 83–93. [CrossRef]
25. Braun, V.; Clarke, V. Using Thematic Analysis in Psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101. [CrossRef]
26. Klarin, A. How to Conduct a Bibliometric Content Analysis: Guidelines and Contributions of Content Co-Occurrence or Co-Word Literature Reviews. *Int. J. Consum. Stud.* **2024**, *48*, e13031. [CrossRef]
27. Wang, Z.Y.; Li, G.; Li, C.Y.; et al. Research on the Semantic-Based Co-Word Analysis. *Scientometrics* **2012**, *90*, 855–875. [CrossRef]
28. Donthu, N.; Kumar, S.; Mukherjee, D.; et al. How to Conduct a Bibliometric Analysis: An Overview and Guidelines. *J. Bus. Res.* **2021**, *133*, 285–296. [CrossRef]
29. Zupic, I.; Čater, T. Bibliometric Methods in Management and Organization. *Organ. Res. Methods* **2015**, *18*, 429–472. [CrossRef]
30. Nguyen, P.M.B.; Pham, X.L.; Truong, G.N.T. A Bibliometric Analysis of Research on Tourism Content Marketing: Background Knowledge and Thematic Evolution. *Heliyon* **2023**, *9*, e13487. [CrossRef]
31. Sharma, K.; Khurana, P. Growth and Dynamics of Econophysics: A Bibliometric and Network Analysis. *Scientometrics* **2021**, *126*, 4417–4436. [CrossRef]

32. Hassan, W.; Duarte, A.E. Bibliometric Analysis: A Few Suggestions. *Curr. Probl. Cardiol.* **2024**, *49*, 102640. [CrossRef]
33. Gyau, E.B.; Sakuwuda, K.; Asimeng, E. A Comprehensive Bibliometric Analysis and Visualization of Publications on Environmental Innovation. *J. Scientometr. Res.* **2023**, *12*, 544–557. [CrossRef]
34. You, C.; Awang, R.; Wu, Y.; et al. Bibliometric Analysis of Global Research Trends on Higher Education Leadership Development Using Scopus Database from 2013–2023. *Discov. Sustain.* **2024**, *5*, 246. [CrossRef]
35. Belmonte, J.L.; Segura-Robles, A.; Moreno-Guerrero, A.J.; et al. Machine Learning and Big Data in the Impact Literature: A Bibliometric Review with Scientific Mapping in Web of Science. *Symmetry* **2020**, *12*, 495. [CrossRef]
36. Alkhammash, R. Bibliometric, Network, and Thematic Mapping Analyses of Metaphor and Discourse in COVID-19 Publications from 2020 to 2022. *Front. Psychol.* **2022**, *13*, 1062943. [CrossRef]
37. Pranckutė, R. Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. *Publications* **2021**, *9*, 12. [CrossRef]
38. Santamaria-Granados, L.; Mendoza-Moreno, J.F.; Ramirez-Gonzalez, G. Tourist Recommender Systems Based on Emotion Recognition—A Scientometric Review. *Future Internet* **2020**, *13*, 2. [CrossRef]
39. Sweileh, W.M. Bibliometric Analysis of Global Scientific Literature on Vaccine Hesitancy in Peer-Reviewed Journals (1990–2019). *BMC Public Health* **2020**, *20*, 1252. [CrossRef]
40. Abdullah, K.H. Publication Trends in Biology Education: A Bibliometric Review of 63 Years. *J. Turk. Sci. Educ.* **2022**, *19*, 465–480. [CrossRef]
41. Yang, Q.; Yang, D.; Li, P.; et al. Resilient City: A Bibliometric Analysis and Visualization. *Discrete Dyn. Nat. Soc.* **2021**, 2021, 5558497. [CrossRef]
42. Yudistira, R.; Rafiek, M.; Herdiani, R.; et al. A Bibliometric Analysis of Sociolinguistic Research in the Past Decade: Trends, Challenges, and Opportunities. *AIP Conf. Proc.* **2022**, *3065*, 030026. [CrossRef]
43. Nurhuda, N.; Gazali, N.; Abdullah, K.H.; et al. Retrospective of Five Years Research of School Leadership in Asia (2018–2022): A Scientometric Paradigm. *Int. J. Eval. Res. Educ.* **2023**, *12*, 1390–1398. [CrossRef]
44. Hamel, R.E. The Dominance of English in the International Scientific Periodical Literature and the Future of Language Use in Science. *AILA Rev.* **2007**, *20*, 53–71. [CrossRef]
45. Lunny, C.; Pieper, D.; Thabet, P.; et al. Managing Overlap of Primary Study Results across Systematic Reviews: Practical Considerations for Authors of Overviews of Reviews. *BMC Med. Res. Methodol.* **2021**, *21*, 140. [CrossRef]
46. Rogers, G.; Szomszor, M.; Adams, J. Sample Size in Bibliometric Analysis. *Scientometrics* **2020**, *125*, 777–794. [CrossRef]
47. McKeown, S.; Mir, Z.M. Considerations for Conducting Systematic Reviews: Evaluating the Performance of Different Methods for De-Duplicating References. *Syst. Rev.* **2021**, *10*, 38. [CrossRef]
48. Goel, A.; Prabha, C.; Sharma, P.; et al. Emerging Research Trends in Data Deduplication: A Bibliometric Analysis from 2010 to 2023. *Arch. Comput. Methods Eng.* **2024**, *31*, 3313–3330. [CrossRef]
49. Hammer, B.; Virgili, E.; Bilotta, F. Evidence-Based Literature Review: De-Duplication a Cornerstone for Quality. *World J. Methodol.* **2023**, *13*, 390–398. [CrossRef]
50. Ruiz-Rosero, J.; Ramirez-Gonzalez, G.; Viveros-Delgado, J. Software Survey: ScientoPy, a Scientometric Tool for Topics Trend Analysis in Scientific Publications. *Scientometrics* **2019**, *121*, 1165–1188. [CrossRef]
51. González-Valiente, C.L.; Costas, R.; Noyons, E.; et al. Terminological (di) Similarities between Information Management and Knowledge Management: A Term Co-Occurrence Analysis. *Mob. Netw. Appl.* **2021**, *26*, 336–346. [CrossRef]
52. Gazali, N.; Saad, N. Job Satisfaction Among Physical Education Teachers: A Scientometric Review. *ASM Sci. J.* **2024**, *19*, 1–13. [CrossRef]
53. van Eck, N.J.; Waltman, L. Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping. *Scientometrics* **2010**, *84*, 523–538. [CrossRef]
54. Cascella, M.; Perri, F.; Ottaiano, A.; et al. Trends in Research on Artificial Intelligence in Anesthesia: A VOSviewer-Based Bibliometric Analysis. *Intelig. Artif.* **2022**, *25*, 126–137. [CrossRef]
55. Kumpulainen, M.; Seppänen, M. Combining Web of Science and Scopus Datasets in Citation-Based Literature Study. *Scientometrics* **2022**, *127*, 5613–5631. [CrossRef]
56. Olmeda-Gómez, C.; Ovalle-Perandones, M.A.; Perianes-Rodríguez, A. Co-Word Analysis and Thematic Landscapes in Spanish Information Science Literature, 1985–2014. *Scientometrics* **2017**, *113*, 195–217. [CrossRef]
57. Howley, I.; Penstein Rosé, C. Modeling the Rhetoric of Human-Computer Interaction. In *Human-Computer*

Interaction. Interaction Techniques and Environments (HCI 2011); Springer: Berlin, Germany, 2011; pp. 341–350. [CrossRef]
58. Morales Sánchez, D.; Moreno, A.; Jiménez López, M.D. A White-Box Sociolinguistic Model for Gender Detection. *Appl. Sci.* **2022**, *12*, 2676. [CrossRef]
59. Grieve, J.; Bartl, S.; Fuoli, M.; et al. The Sociolinguistic Foundations of Language Modeling. *Front. Artif. Intell.* **2025**, *7*, 1472411. [CrossRef]
60. Abitbol, J.L.; Karsai, M.; Magué, J.P.; et al. Socioeconomic Dependencies of Linguistic Patterns in Twitter: A Multivariate Analysis. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 1125–1134. [CrossRef]
61. Tarrade, L.; Magué, J.P.; Chevrot, J.P. Detecting and Categorising Lexical Innovations in a Corpus of Tweets. *Psychol. Lang. Commun.* **2022**, *26*, 313–329. [CrossRef]
62. Nissan, E. ONOMATURGE: An Artificial Intelligence Tool and Paradigm for Supporting National and Native Language Fostering Policies. *AI Soc.* **1991**, *5*, 202–217. [CrossRef]
63. Hovy, D.; Rahimi, A.; Baldwin, T.; et al. Visualizing Regional Language Variation across Europe on Twitter. In *Handbook of the Changing World Language Map*; Springer: Cham, Switzerland, 2019; pp. 3719–3742. [CrossRef]
64. Mengesha, Z.; Heldreth, C.; Lahav, M.; et al. "I Don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Front. Artif. Intell.* **2021**, *4*, 725911. [CrossRef]
65. Astuti, L.W.; Sari, Y. Code-Mixed Sentiment Analysis Using Transformer for Twitter Social Media Data. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 498–504. [CrossRef]
66. Guo, Z.; Lai, A.; Thygesen, J.H.; et al. Large Language Models for Mental Health Applications: Systematic Review. *JMIR Ment. Health* **2024**, *11*, e57400. [CrossRef]
67. Waliya, Y.J. Technolingualism and Multilingualism on the Web 3.0: Ubek Et Rica Blog. *Ezikov Svyat (Orbis Linguarum)* **2024**, *22*, 118–130. [CrossRef]
68. Tran, H.; Stell, A. Beyond Borders or Building New Walls?: The Potential for Generative AI in Recolonising the Learning of Vietnamese Dialects and Mandarin Varieties. *Aust. Rev. Appl. Linguist.* **2024**, *47*, 284–308. [CrossRef]
69. Xiao, Y.; Yu, S. Can ChatGPT Replace Humans in Crisis Communication? The Effects of AI-Mediated Crisis Communication on Stakeholder Satisfaction and Responsibility Attribution. *Int. J. Inf. Manage.* **2025**, *80*, 102835. [CrossRef]
70. Yibokou, K.S.; Boulton, A.; Kalyaniwala, C.; et al. Spontaneous Use of Generative Artificial Intelligence and Influence on Collaborative Learner Writing. *Alsic* **2025**, *28*. [CrossRef]
71. Zhang, X.; Umeanowai, K.O. Exploring the transformative influence of artificial intelligence in EFL context: A comprehensive bibliometric analysis. *Educ. Inf. Technol.* **2025**, *30*, 3183–3198. [CrossRef]
72. Blommaert, J. Sociolinguistic Restratification in the Online-Offline Nexus: Trump's Viral Errors. In *Language Policies and the Politics of Language Practices*; Springer: Cham, Switzerland, 2021; pp. 7–24. [CrossRef]
73. Sanei, T. Normativity, Power, and Agency: On the Chronotopic Organization of Orthographic Conventions on Social Media. *Lang. Soc.* **2022**, *51*, 453–480. [CrossRef]
74. Laitinen, M.; Fatemi, M.; Lundberg, J. Size Matters: Digital Social Networks and Language Change. *Front. Artif. Intell.* **2020**, *3*, 46. [CrossRef]
75. Laitinen, M.; Lundberg, J. ELF, Language Change, and Social Networks: Evidence from Real-Time Social Media Data. In *Language Change: The Impact of English as a Lingua Franca*; Cambridge University Press: Cambridge, UK, 2020; pp. 179–204. [CrossRef]
76. Goel, R.; Soni, S.; Goyal, N.; et al. The Social Dynamics of Language Change in Online Networks. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2016; pp. 41–57. [CrossRef]
77. Gonzales, W.D.W. When to (not) Split the Infinitive: Factors Governing Patterns of Syntactic Variation in Twitter-Style Philippine English. *Eng. Lang. Linguist.* **2024**, *28*, 305–339. [CrossRef]
78. Grondelaers, S.; Van Hout, R.; Van Halteren, H.; et al. Why Do We Say Them When We Know it Should be They? Twitter as a Resource for Investigating Nonstandard Syntactic Variation in the Netherlands. *Lang. Var. Change* **2023**, *35*, 223–245. [CrossRef]
79. Konisi, L.Y.; Aso, L.; Taembo, M. The Maintenance of Landawe Language and Its Correlation to People's Attitudes in North Konawe, Southeast Sulawesi. *Linguistica Silesiana* **2024**, *45*, 135–152.
80. Dijkstra, J.; Heeringa, W.; Jongbloed-Faber, L.; et al. Using Twitter Data for the Study of Language Change in Low-Resource Languages. A Panel Study of Relative Pronouns in Frisian. *Front. Artif. Intell.* **2021**, *4*, 644554.

[CrossRef]
81. Alshaabi, T.; Dewhurst, D.R.; Minot, J.R.; et al. The Growing Amplification of Social Media: Measuring Temporal and Social Contagion Dynamics for over 150 Languages on Twitter for 2009–2020. *EPJ Data Sci.* **2021**, *10*, 15. [CrossRef]
82. Krishna, D.; Gupta, V.; Kumari, K.; et al. Impact of AI-Powered Translation Tools. *J. Digit. Sociohumanit.* **2025**, *2*, 70–80. [CrossRef]
83. Park, S. AI Chatbots and Linguistic Injustice. *J. Univ. Lang.* **2024**, *25*, 99–119. [CrossRef]
84. Knauth, J. Language-Agnostic Twitter-Bot Detection. In Proceedings of the Recent Advances in Natural Language Processing, Varna, Bulgaria, 2–4 September 2019; pp. 550–558. [CrossRef]
85. Simaki, V.; Mporas, I.; Megalooikonomou, V. Age Identification of Twitter Users: Classification Methods and Sociolinguistic Analysis. In Proceedings of the 17th International Conference, CICLing 2016, Konya, Turkey, 3–9 April 2018; pp. 385–395. [CrossRef]
86. Devi, V.; Sharma, A. Sentiment Analysis Approaches, Types, Challenges, and Applications: An Exploratory Analysis. In Proceedings of the 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, India, 25–27 November 2022; pp. 34–38. [CrossRef]
87. Barakat, A.; Al Hammadi, O.; Aldhaheri, A.; et al. Arabic Dialect Identification from Speech. In Proceedings of the 2024 15th Annual Undergraduate Research Conference on Applied Computing (URC), Dubai, United Arab Emirates, 24–25 April 2024. [CrossRef]
88. Glazkova, A.; Egorov, Y.; Glazkov, M. A Comparative Study of Feature Types for Age-Based Text Classification. In *Analysis of Images, Social Networks and Texts*; Springer: Cham, Switzerland, 2021; pp. 120–134. [CrossRef]
89. Hagos, D.H.; Battle, R.; Rawat, D.B. Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives. *IEEE Trans. Artif. Intell.* **2024**, *5*, 5873–5893. [CrossRef]
90. Habernal, I.; Gurevych, I. Argumentation Mining in User-Generated Web Discourse. *Comput. Linguist.* **2017**, *43*, 125–179. [CrossRef]
91. Strzalkowski, T.; Shaikh, S.; Liu, T.; et al. Influence and Power in Group Interactions. In *Social Computing, Behavioral-Cultural Modeling and Prediction*; Springer: Berlin, Germany, 2013; pp. 19–27. [CrossRef]
92. Tikhonova, E.; Raitskaya, L. ChatGPT: Where Is a Silver Lining? Exploring the Realm of GPT and Large Language Models. *J. Lang. Educ.* **2023**, *9*, 5–11. [CrossRef]
93. Dergaa, I.; Chamari, K.; Zmijewski, P.; et al. From Human Writing to Artificial Intelligence Generated Text: Examining the Prospects and Potential Threats of ChatGPT in Academic Writing. *Biol. Sport* **2023**, *40*, 615–622. [CrossRef]
94. Lund, B.; Wang, T.; Mannuru, N.R.; et al. ChatGPT and a New Academic Reality: Artificial Intelligence-Written Research Papers and the Ethics of the Large Language Models in Scholarly Publishing. *J. Assoc. Inf. Sci. Technol.* **2023**, *74*, 570–581. [CrossRef]
95. Lepp, H.; Smith, D.S. "You Cannot Sound Like GPT": Signs of Language Discrimination and Resistance in Computer Science Publishing. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, Athens, Greece, 23–26 June 2025; pp. 3162–3181. [CrossRef]
96. Dunn, J.; Edwards-Brown, L. *Geographically-Informed Language Identification*; OSF: Charlottesville, VA, USA, 2024. [CrossRef]
97. Doval, Y.; Vilares, M.; Vilares, J. On the Performance of Phonetic Algorithms in Microtext Normalization. *Expert Syst. Appl.* **2024**, *113*, 213–222. [CrossRef]
98. Curry, A.C.; Attanasio, G.; Talat, Z.; et al. Classist Tools: Social Class Correlates with Performance in NLP. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, August 2024; pp. 12643–12655. [CrossRef]