Intelligent Systems with Applications 24 (2024) 200456

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Review

Ideological orientation and extremism detection in online social networking sites: A systematic review

Kamalakkannan Ravi [*], Jiann-Shiun Yuan

Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816, USA

ARTICLE INFO

Keywords:
Extremism detection
Machine learning
Natural language processing
Predictive models
Social media
User-generated content

ABSTRACT

The rise of social networking sites has reshaped digital interactions, becoming fertile grounds for extremist ideologies, notably in the United States. Despite previous research, understanding and tackling online ideological extremism remains challenging. In this context, we conduct a systematic literature review to comprehensively analyze existing research and offer insights for both researchers and policymakers. Spanning from 2005 to 2023, our review includes 110 primary research articles across platforms like Twitter (X), Facebook, Reddit, TikTok, Telegram, and Parler. We observe a diverse array of methodologies, including natural language processing (NLP), machine learning (ML), deep learning (DL), graph-based methods, dictionary-based methods, and statistical approaches. Through synthesis, we aim to advance understanding and provide actionable recommendations for combating ideological extremism effectively on online social networking sites.

Contents

[*] Corresponding author.
E-mail addresses: kamalakkannan.ravi@ucf.edu (K. Ravi), jiann-shiun.yuan@ucf.edu (J.-S. Yuan).

https://doi.org/10.1016/j.iswa.2024.200456
Received 22 August 2024; Received in revised form 21 October 2024; Accepted 28 October 2024
Available online 8 November 2024

1. Introduction

The advent of the internet in the early 1990s, catalyzed by pioneers like Berners-Lee (Gillies & Cailliau, 2000), marked a profound shift in global communication and connectivity. What started as rudimentary web pages outlining the World Wide Web project has since evolved into a vast digital ecosystem, enabling individuals worldwide to engage in uninhibited exchanges of ideas, opinions, and ideologies. This digital transformation has brought the proliferation of online platforms, ranging from social networking sites to forums and blogs, providing users unprecedented avenues for connection and self-expression.

Yet, alongside these opportunities for discourse and expression, the online sphere has increasingly become a fertile ground for the propagation of extremist ideologies and activities. Dedicated social networking platforms such as Classmates, SixDegrees, Friendster, MySpace, Facebook, Orkut, Reddit, Twitter (for consistency throughout this paper, the platform X will be referred to by its common name, Twitter), Tumblr, Pinterest, Instagram, Quora, Snapchat, Google+, Twitch, Telegram, Vine, Discord, TikTok, and Clubhouse, among others, have provided platforms for individuals and groups espousing extremist views to connect, organize, and disseminate their ideologies (Holt, Freilich, & Chermak, 2022; Liang, 2022; O'Hara & Stevens, 2015; Walther & McCoy, 2021). From neo-Nazi and alt-right activists to occupy movement and far-right extremists, these actors leverage social media networks to propagate their beliefs, recruit followers, and mobilize support (Fernandez, Asif, & Alani, 2018; Walther & McCoy, 2021).

The impact of social media in enabling the proliferation of extremist groups has been particularly pronounced in the United States (Erbschloe, 2018), where political polarization and ideological schisms have become increasingly salient. The rise of online platforms has amplified these divisions, with users often retreating into echo chambers where their convictions are reinforced, and dissenting voices are marginalized (O'Hara & Stevens, 2015). This phenomenon has been further exacerbated by foreign entities, exemplified by the Russian Internet Research Agency, which exploits social media platforms to agitate discord and manipulate public opinion, as evidenced in the 2016 United States presidential election (Department of Justice, 2018; Linvill, Boatwright, Grant, & Warren, 2019). Similar tactics have been observed in other geopolitical contexts, such as public reactions to the Russo–Ukraine War, where social media, particularly Twitter, has been used to shape narratives and influence public sentiment globally (Tamer et al., 2023), and where media influence on public opinion continues to be significant (Zaytoon, Bashar, Khamis, & Gomaa, 2024).

Early efforts by researchers like Davidson, Warmsley, Macy, and Weber (2017) and Zampieri et al. (2019) laid the groundwork for extremism detection by categorizing social media posts into hate speech, offensive language, and neutral content. Melton, Bagavathi, and Krishnan (2020)advanced this with the Offensive Language Identification Dataset (OLID), refining hate speech classification, while Waseem (2016) contributed by labeling Twitter posts as Racist, Sexist, or Harmless, enhancing the granularity of harmful language detection. These foundational studies were crucial in identifying extremism and ideological polarization using various models. Well-labeled datasets have also been essential for training effective models, such as Waseem's dataset with 15,216 tweets, Davidson et al.'s Hate or Offensive (HON) dataset with around 25,000 tweets, and OLID with 13,000 tweets labeled for hate speech. Domain-specific datasets, like Behzadan et al.'s 21,000 cyber-related tweets (Behzadan, Aguirre, Bose, & Hsu, 2018) and Mahata et al.'s (Mahata et al., 2019) work on targeted offenses, along with YouTube-specific datasets, further expand the resources available for detecting extremism across platforms.

While these studies have made significant contributions, there is a lack of a unified review that integrates insights across these varied efforts. A comprehensive synthesis of different datasets, methodologies, and platforms is needed to provide a clearer understanding of the state-of-the-art in extremism detection. Additionally, a temporal analysis is necessary to trace the evolution of detection techniques and assess their real-world applications. Lastly, a focused examination of ongoing challenges and future research opportunities is essential to guide innovation in the field. This review addresses these gaps by offering a systematic analysis of existing literature, identifying key trends, and providing actionable recommendations for researchers and policymakers aiming to combat ideological extremism in online spaces, particularly in regions like the United States, where ideological polarization is a growing concern (Davis, 2017).

Through this lens, the review highlights three core research gaps: the need for a comprehensive synthesis of existing efforts, the importance of analyzing the historical and practical evolution of detection methods, and the ongoing challenges that require future research. Our systematic literature review seeks to fill these gaps and serve as a vital resource for advancing the study of ideological extremism on social media.

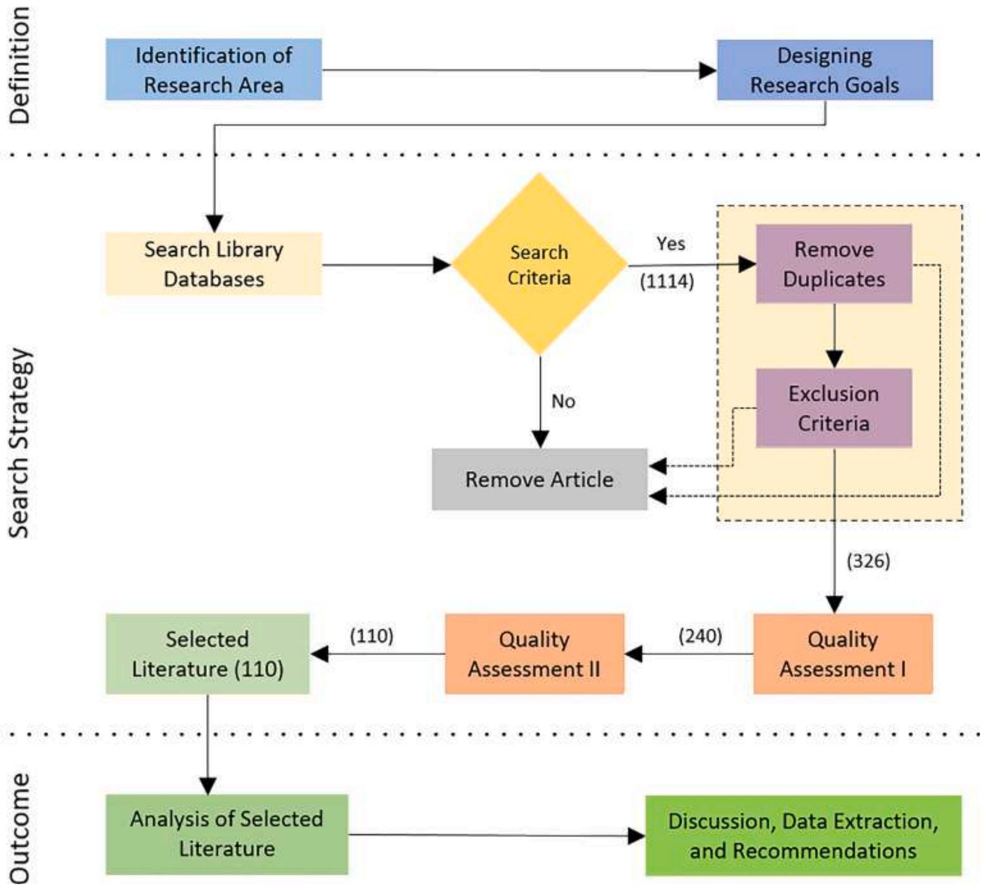

Fig. 1. Search strategy based on PRISMA guidelines.

Research Gap 1: Comprehensive Synthesis. Existing studies have explored various aspects of ideological extremism detection, yet there remains a notable absence of a comprehensive synthesis that integrates insights across diverse research domains. This synthesis is vital for clarifying state-of-the-art techniques.

Research Gap 2: Temporal Analysis and Practical Applications. Understanding the historical evolution of ideological and extremism detection techniques and their practical applications is crucial for assessing their real-world impact. However, there is a lack of research focusing on temporal analysis and translating advancements into practical solutions.

Research Gap 3: Challenges and Recommendations. Despite the progress made in ideological extremism detection, numerous challenges persist, and several areas remain unexplored or underexplored. Moreover, identifying potential avenues for future research and delineating the advancements is essential for continued innovation. However, there is a lack of comprehensive analyses that systematically identify challenges, highlight unexplored areas, and offer recommendations for future research directions.

By addressing these research gaps, our systematic literature review aims to provide valuable insights for researchers and policymakers, guiding efforts to effectively combat ideological extremism in online social networking sites.

2. Methodology

In this systematic literature review, we adopt a structured approach to gather and analyze pertinent literature (Brereton, Kitchenham, Budgen, Turner, & Khalil, 2007; Denyer & Tranfield, 2009; Keele et al., 2007; Petersen, Vakkalanka, & Kuzniarz, 2015; Sarkar, 2022; Tranfield, Denyer, & Smart, 2003). Drawing from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Tricco et al., 2018), databases were systematically searched for articles. Our methodology involves adapting systematic review methodologies to align with the specific requirements of our research objectives. We outline the key steps involved in this process, beginning with the clear definition of research questions and culminating in the synthesis of findings.

To ensure comprehensiveness, we employ strategies for conducting thorough literature searches, encompassing diverse sources and search terms (Cortis & Davis, 2021). Through this process, we aim to identify key research themes, methodologies, and findings from existing studies in the field. By systematically reviewing the literature, we endeavor to provide a comprehensive understanding of the subject matter and contribute to the advancement of knowledge in the research domain.

We delve into the specifics of our review process, delineating each step to provide transparency and clarity, as shown in Fig. 1. The subsections are structured as follows:

I. Definition: Research area and questions
II. Search strategy

(a) Databases
(b) Selection of indexed search terms
(c) Search application
(d) Study selection

III. Outcome: Discussion, data extraction, and recommendations

2.1. *Definition: Research area and questions*

We establish the research area and frame questions to guide our inquiry. Drawing from the backdrop of increasing ideological polarization and extremism on social media, notably pronounced within

Table 1
Selected electronic library databases.

RQ1	1. Computer Science Database (ProQuest) (1981+)
	2. Applied Science & Technology Source (EBSCOhost) 3. Web of Science (1965+) 4. Compendex (Ei Engineering Village) (1884+)
RQ2	1. Compendex (Ei Engineering Village) (1884+)

the politically divided landscape of the United States, our review aims to explore existing research on ideological orientation and extremism detection in online social networking sites.

Next, we define research questions to bridge the research gaps identified earlier in the Introduction Section 1, serving as the foundational framework for our investigation. Our focus is on synthesizing research articles related to the use of machine learning (ML), natural language processing (NLP), deep learning (DL), graph-based methods, dictionary-based methods, and statistical techniques for identifying ideological orientation and extremism. Following established review protocols, we formulate the following research questions:

RQ1 What methods are employed for detecting ideological orientation, particularly on social networking sites in the United States, with a focus on discerning differences between liberal and conservative, right and left, or Democrat and Republican ideologies?

RQ2 What is the state of machine learning and natural language processing techniques in identifying ideological extremism, especially within social networking sites in the United States?

2.2. Search strategy

Our search strategy for this systematic review primarily targets published papers, including journals, conference and workshop proceedings, and book chapters.

2.2.1. Databases

We curated a selection of databases to ensure comprehensive coverage of relevant scholarly sources. Our chosen databases include ProQuest,[1] EBSCO Host,[2] Web of Science,[3] and Ei Compendex.[4] Detailed information regarding these databases for each research question is presented in Table 1. During our literature collection process, we observed that Ei Compendex stands out as a particularly exhaustive resource. It serves as a comprehensive database for accessing articles related to ML, NLP, DL, graph-based methods, dictionary-based methods, and statistical methods for ideological orientation and extremism detection. Consequently, we chose to exclusively utilize this database for research question 2, as indicated in Table 1.

2.2.2. Selection of indexed search terms

Search terms were systematically selected to ensure comprehensive coverage of relevant literature aligned with our research questions. Specifically, these terms were chosen to address ideological affiliation and ideological extremism, as outlined in previous studies (Simons & Skillicorn, 2020; Walther & McCoy, 2021). We selected these seed terms and expanded upon them during the search process by selecting relevant index terms within the databases listed in Table 1 during literature collection. Table 2 presents the indexed search terms used in our study.

2.2.3. Search application

Our search strategy involved applying selected search terms within designated databases using Boolean operators. The results were further refined through the following inclusion and exclusion criteria:

- Inclusion Criteria:
  - Articles focusing on the United States.
  - Articles written in English.
  - Articles discussing topics related to politics or social networking.
  - Articles utilizing ML, NLP, DL, graph-based methods, dictionary-based methods, and statistical methods for detecting ideological orientation and extremism.
- Exclusion Criteria:
  - Articles published in languages other than English.
  - Papers presenting secondary studies.
  - Non-conference or non-journal papers, such as reviews, theses, dissertations, and magazines.
  - Articles that utilized ML, NLP, DL, graph-based methods, dictionary-based methods, and statistical techniques without detecting ideological orientation and extremism.
  - Articles deemed irrelevant based on the title, abstract, and conclusion.

2.3. Study selection

The study selection process began with retrieving 1114 studies meeting the inclusion criteria. After removing duplicates and applying exclusion criteria, 326 studies remained for further assessment. Among these, 253 articles were identified for research question 1, sourced from ProQuest (48), EBSCOhost (41), EICompendex (124), and Web of Science (40) databases. For research question 2, 73 articles were retrieved from EICompendex. Mendeley,[5] an open-source tool, facilitated reading, annotation, and categorization of papers according to our research questions.

Subsequently, we conducted the first round of quality assessment (QA) by scrutinizing article content for relevance to research questions 1 and 2. Snowballing was employed to include relevant literature, resulting in the retention of 240 articles (138 for RQ1 and 102 for RQ2). In the subsequent QA round, we assessed whether these articles utilized ML, NLP, or survey techniques to identify ideological orientation and extremism. Snowballing was again used to include pertinent literature, leading to the identification of 110 papers (69 for RQ1 and 41 for RQ2) meeting our criteria.

Consequently, a total of 110 primary studies from conferences and journals published between 2005 and 2023 were identified, as depicted in Fig. 2.

3. Findings RQ1: Ideological orientation detection in the U.S

This section delves into various research papers that have investigated ML, NLP, DL, graph-based methods, dictionary-based methods, and statistical approaches for identifying political orientation. The objective is to understand and differentiate political affiliations, especially on social networking sites in the United States. The emphasis lies in discerning disparities between liberal and conservative, right and left, or Democrat and Republican ideologies. We have organized these research papers under the following themes to address our research question RQ1.

[1] https://www.proquest.com/
[2] https://search.ebscohost.com/
[3] https://www.webofscience.com/
[4] https://www.engineeringvillage.com/
[5] https://mendeley.com/

Table 2
Search terms.

RQ1	social media, sentiment analysis, popular vote, propaganda, polarization microblogs, fake news, politics left wing and right wing, united states, conservative and liberal, social networking (online), democrat and republican, conservative bias, liberal bias, populism, nationalism
RQ2	social networking (online), american nazi party, christianity, white supremacists, extremism, the far right, radical politics, hate groups, alternative right, conservative, disinformation, politics, learning (artificial intelligence), hate crimes, facebook, radical, immigration, white supremacists, zionism, anti-semitism, christian, pro-life, abortion, human rights, anarchy, politics, eugenics, united states, politics, tea party, politics, immigration, fascists, right wing extremism, political campaigns

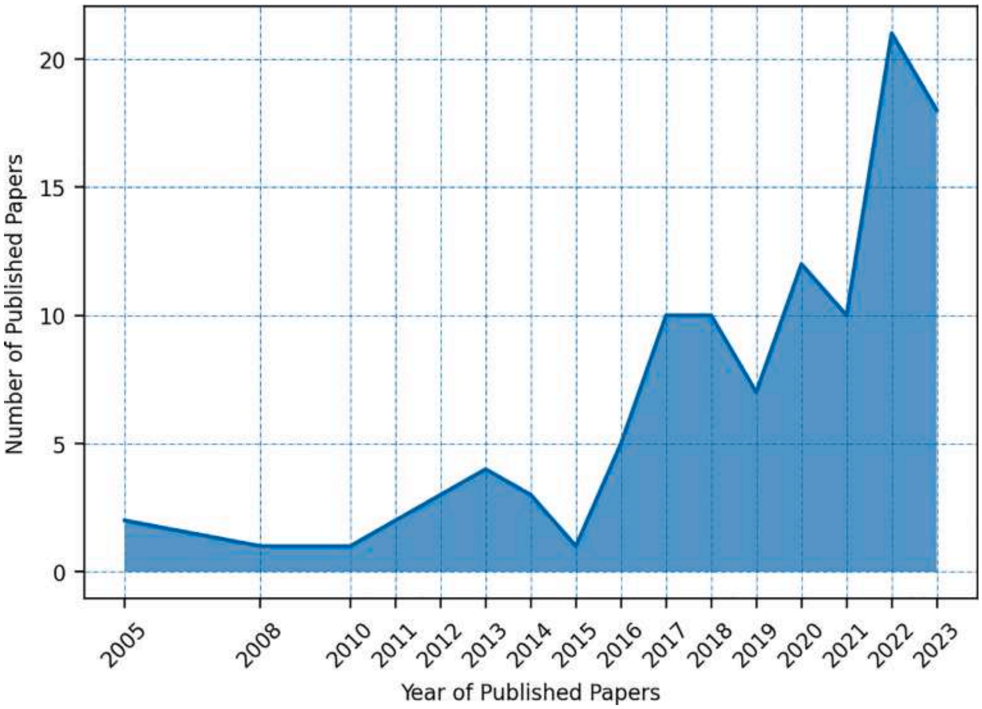

Fig. 2. Primary Studies by Year.

3.1. Political ideological analysis on social media

In the exploration of ideological analysis on social media, Golbeck and Hansen (2011) examined methods to estimate political preferences on Twitter, particularly focusing on U.S. media outlets. By utilizing follower connections, they were able to gauge liberal/conservative scores from Congress members to media outlet audiences. Their findings revealed a close match between the political leanings of media outlets and the preferences of their audiences, showcasing the potential of text mining for personalization and bias detection in social media. Similarly, Himelboim, McCreery, and Smith (2013) delved into how Twitter users interact with cross-ideological political content. Through network and content analyses, they mapped Twitter networks on contentious political topics, revealing that users tend to stick to ideological bubbles, with liberal viewpoints mainly tied to traditional media sources. This trend was further highlighted by Colleoni, Rozza, and Arvidsson (2014) when they classified Twitter users as Democrats or Republicans using machine learning and social network analysis based on their political content sharing. They discovered a higher political homophily among Democrats in their online networks, emphasizing the importance of considering users' political behaviors when studying political homophily on social media. Additionally, Lahoti, Garimella, and Gionis (2018) proposed a machine-learning approach to model the liberal-conservative ideology space, aiming to identify ideological leaning for users and media sources, addressing the issue of information filter bubbles prevalent in the digital information landscape.

On exploring the 2012 U.S. presidential election with respect to the partisans and political actors, Stier (2016) examined partisan framing in Twitter debates during the 2015 U.S. political discussions. Utilizing framing theory and computational text analysis, they identified semantic differences between tweets from Democratic and Republican actors, shedding light on the strategies used to shape political narratives on social media. Similarly, Noel (2016) examined the dynamics within political parties, focusing on the interplay between party regulars and ideologues. Their findings revealed a notable difference: Democrats exhibited a less pronounced division, with party regulars predominantly guiding the party, whereas Republicans faced a deeper internal divide, influencing their strategies and presenting challenges during primary elections. Further analyzing the complex dynamics between candidate posts and commentator sentiment on Facebook during the election, Alashri et al. (2016) demonstrated the intricate correlations between online sentiment and offline events, as well as candidate strategies.

On the other hand, Wong, Tan, Sen, and Chiang (2016) tackled the political orientation of Twitter with respect to users during the election, considering tweet and retweet patterns. Their study provided valuable insights into the political makeup of the Twitter population and the evolving dynamics of political polarization over time. Continuing this trajectory, Preoţiuc-Pietro, Liu, Hopkins, and Ungar (2017) characterized politically engaged users based on their language use on Twitter and ventured into predicting political ideology on a seven-point scale. Their research contributed significantly to a nuanced understanding of the multifaceted and nuanced political orientations that permeate the world of social media.

3.2. Advanced techniques for political ideological analysis

In this section, we explore cutting-edge methods to delve into the complexities of digital political discourse. Malouf and Mullen (2008) extended natural language processing techniques to informal online political discussions, emphasizing the importance of incorporating social network analysis to enhance classification accuracy. As digital platforms evolved, researchers adapted their analytical techniques. On supervised learning, Alzhrani (2022) (2022) has ventured into the realm of automating the detection of two pivotal political science concepts: politician personalization and political ideology. This research demonstrates a significant improvement in political ideology detection models by incorporating deep neural network models based on politicians' personalization. This fusion of techniques holds the potential to revolutionize our understanding of the subtle nuances within political ideologies as expressed in digital spaces. On the other hand, Fagni and Cresci (2022) introduced an innovative unsupervised deep learning approach to predict the political leaning of social media users. By harnessing deep neural networks and clustering techniques, they achieved fine-grained predictions of political orientation, offering a nuanced understanding of user ideologies in the digital age.

Moreover, Pennacchiotti and Popescu (2011) proposed a machine learning framework for automatically constructing user profiles on Twitter, focusing on classifying users based on linguistic content, social behaviors, and social graph information. Tasks include detecting political affiliation, identifying ethnicity, and determining affinity for businesses like Starbucks. Chiu and Hsu (2018) concentrated on examining left and right-wing political posts on Facebook in the U.S. They aim to predict the political orientation (left or right) based on the sentiment of these posts. Using sentiment analysis with lexical databases to detect emotional words, the study evaluates different classification algorithms and assesses prediction performance.

In the exploration of online political discourse, researchers have investigated diverse aspects, including the usage of political slang, interpretation of ideological scales, and communication patterns on emerging platforms like TikTok. Hossain, Tran, and Kautz (2018) explore innovative political slang in the comment sections of online political news articles during the 2016 US Presidential Election. They define creative political slang and devise an unsupervised algorithm, PoliSlang, to identify such slang in reader comments. The study aims to compare and contrast political slang usage across various news media platforms with differing political inclinations. Simas (2018) examines the interpretation of the 7-point ideological scale in U.S. surveys, particularly the terms "liberal" and "conservative". Introducing anchoring vignettes, the study links interpretation differences to partisanship. Democrats and Republicans interpret the scale differently, with Democrats having lower thresholds for distinctions between categories. The paper suggests that neglecting these perceptual differences may underestimate the ideological gap between the two parties. Medina Serrano, Papakyriakopoulos, and Hegelich (2020) examine political communication on TikTok in their paper "Dancing to the Partisan Beat: A First Analysis of Political Communication on TikTok". The study delves into partisan Republican and Democratic videos in the U.S., utilizing computer vision, natural language processing, and statistical tools to analyze political communication on the platform. It explores user demographics, interaction structures, and the nature of political discourse on TikTok.

Additionally, Tien, Eisenberg, Cherng, and Porter (2020) conducted a case study on the Twitter conversation post the 2017 'Unite the Right' rally in Charlottesville, Virginia, USA. The study employs network analysis and data science tools to investigate the online conversation's structure, focusing on polarization, influential accounts, and community structures. The authors classify Twitter accounts based on media preferences and determine a 'Left/Right' orientation score. Findings reveal a highly polarized retweet network, with an analysis of content differences between Left-leaning and Right-leaning communities. Decter-Frain and Barash (2022) investigated the use of knowledge graph embeddings (KGEs) for detecting and analyzing partisanship in online political discourse. Unlike traditional methods that focus on linguistic or network aspects within a single platform, this study explores the potential of KGEs to integrate linguistic and network information across various platforms. The authors employ a semi-supervised approach to reveal a political dimension in the embedding space, demonstrating that KGEs enable more accurate differentiation between liberal and conservative Twitter accounts. They apply this method to discussions on COVID-19 and climate change, showcasing the generalizability of the findings. Ramaciotti Morales (2022) addressed dysfunctions in online social networks, such as echo chambers and filter bubbles, by analyzing user opinions in multidimensional ideological spaces. The paper addresses limitations in studying online political behavior, particularly in the U.S., due to the dominance of a principal liberal-conservative cleavage. The author proposes a method utilizing social graph embedding and natural language processing techniques to identify additional cleavage dimensions related to cultural, policy, social, and ideological groups and preferences. Utilizing Twitter data from nearly 2 million users engaged in the U.S. political debate, the method aims to uncover non-aligned dimensions beyond the traditional liberal-conservative divide.

To explore user-driven engagement, Zerrer and Engelmann (2022) employed the social identity model of collective action (SIMCA) to decipher political motivations expressed through user comments on news sites. By identifying various clusters of user comments based on political motivations, their study shed light on the motivations behind user engagement with news content. In a similar vein, Ravi, Vela, and Ewetz (2022) utilized NLP techniques to classify user-submitted texts on social media, focusing on distinguishing between conservative and liberal content. Through meticulous language analysis and evaluation of classifiers, this research offered a granular understanding of ideological diversity within digital communities. Further exploring Reddit discourse, Botzer and Weninger (2023) introduced the concept of "entity graphs" to analyze online discussions, particularly focusing on discussions from Reddit. Although not explicitly identifying political ideologies, their research delved into the communication patterns of conservative and liberal groups on specific subreddits, revealing insights into affective polarization and online echo chamber formation. Moreover, Alkiek, Zhang, and Jurgens (2022) tackled the intricate task of identifying political users within the diverse landscape of Reddit and explored the implications of different user definitions on modeling and analysis. Their research revealed the complexities of political inference within online communities, shedding light on the intricate dynamics of ideological identification. Continuing the exploration of social media platforms, Olteanu, Cernian, and Gâgă (2022) employed NLP techniques to discern the political orientation of users on Twitter. Leveraging machine learning and semi-structured information, they successfully classified users as Democrat or Republican based on their posts, achieving a remarkably high level of accuracy and opening doors to more precise political ideology detection in the digital realm.

Moreover, Fichman and Akter (2023) conducted a comprehensive analysis of online trolling during the 2016 and 2020 US presidential election cycles, focusing on ideological asymmetry. Their study dissected trolling tactics on Twitter, highlighting notable differences in engagement between Republicans and Democrats. Similarly, Hashemi (2023) employed a multidisciplinary approach, combining deep learning, NLP, geographical information systems (GIS), and statistical tools, to visualize and analyze political misinformation, extremism, and topics on social media during the USA 2020 presidential election. This innovative approach provided valuable insights into the geographical dimensions of digital political discourse. Also, González-Bailón et al. (2023) delved into the phenomenon of ideological segregation in political news consumption on Facebook during the US 2020 election. Their study shed light on the concentration of misinformation within a homogeneously conservative segment of the news ecosystem, revealing the dynamics of ideological polarization.

3.3. Media bias and perception

Commencing with bias exploration, Lee (2005) delved into the perceptions of media bias among the general public and politicians, with a specific focus on the belief in a liberal bias within U.S. news media. Drawing from national surveys, the study explores whether this perception is associated with an observer's partisan and ideological positions. The findings reveal that strong conservatives and Republicans are more inclined to distrust the news media, with political cynicism playing a pivotal role in shaping perceptions of media bias. Concurrently, Adamic and Glance (2005) investigated the linking patterns and discussion topics of political bloggers leading up to the 2004 U.S. Presidential Election. Their study aims to differentiate between liberal and conservative blogs, analyzing posts from 40 "A-list" blogs over two months and a snapshot of over 1000 political blogs in a single day. The research reveals that conservative blogs exhibit denser linking patterns and more frequent interlinking compared to liberal ones. By employing web crawling, data analysis, and network analysis techniques, the study sheds light on the distinct characteristics of the political blogosphere during the election period, particularly focusing on ideological differences in linking behavior and discussion topics. Further, the U.S. political blogs on the left and right were investigated by, Shaw and Benkler (2012) in 2008. They analyze ideological affiliations, technologies, institutions, and participation practices to understand the Internet's impact on democratic practice and knowledge production. From the scrutiny of media outlets to the intricacies of user attention and information dissemination, researchers have delved into diverse aspects of media bias, perception, and their impact on public discourse.

For instance, during the 2012 presidential election, Kaye and Johnson (2016) investigated media consumption patterns and the impact of online sources, partisanship, and perceived media bias during the 2012 presidential election. Their study examined whether online sources influenced traditional media consumption, considering politically neutral and biased sources, while also exploring the role of perceived media bias against political candidates in shaping individuals' time spent with various media outlets. This research provided valuable insights into the intricate relationship between media consumption, partisanship, and perceptions of bias. On the other hand, Le, Shafiq, and Srinivasan (2017) introduced a method for efficiently measuring the political bias of news articles using Twitter data. By analyzing how news articles are shared on Twitter and examining users' connections to key Democrat and Republican accounts, they estimated the ideological slant of articles, validated against crowdsourced slant labels, showcasing accurate classification of Democratic and Republican-leaning news articles. Further, Ribeiro et al. (2018) innovatively introduced the "Media Bias Monitor" methodology, discerning biases in thousands of news sources on social media by scrutinizing audience demographics. Their work underscores the critical importance of evaluating biases within news outlets and highlights how these biases profoundly impact the information landscape.

The emergence of social media as a dominant news consumption platform has heightened concerns about bias and misinformation. Chen, Pacheco, Yang, and Menczer (2021) shifted focus towards biases in news and information encountered on platforms like Twitter, particularly within the context of U.S. political discourse. Their research deployed "drifter" bots with neutral behavior to probe exposure biases, political echo chambers, and the spread of misinformation, offering insights into the complex landscape of news consumption and information exposure on social media. This study underscores the challenges posed by echo chambers and the propagation of biased content in shaping public discourse. Additionally, Morris, Morris, and Francia (2020)'s post-election study underscored the challenges of combatting misinformation and the role of pre-existing beliefs in susceptibility to false information.

Shifting the focus from user behavior to the assessment of trustworthiness in media consumption, Neo (2021)'s investigation explored the correlation between perceived political network homogeneity on social media, news credibility, and political engagement. Meanwhile, Martel, Allen, Pennycook, and Rand (2022)'s research centered on enhancing the reliability of crowdsourced assessments in evaluating online news, considering various elicitation processes and their impact on judgment quality.

Moving from media outlets to user attention, Morgan, Lampe, and Shafiq (2013) explored the impact of perceived ideological bias in news outlets on the sharing of news content on Twitter. They conducted a meticulous analysis of tweets referencing popular news outlets across the ideological spectrum, uncovering differences in news sharing patterns based on outlets or perceived ideological leanings. This research contributes to our understanding of the selective exposure theory, where individuals tend to consume and share news that aligns with their pre-existing beliefs. Similarly, Mason and Wronski (2018) sought to uncover ideological asymmetries observable on platforms like YouTube and Twitter. Their innovative approach introduced cross-platform metrics aimed at quantifying the dynamics of attention across different ideological groups. This research emphasized the profound impact of social identities in shaping partisan attachments within the intricate tapestry of digital engagement.

Transitioning from user attention to language models, Liu, Jia, Wei, Xu, and Vosoughi (2022) addressed political bias within language models like GPT-2, proposing methodologies to mitigate bias while maintaining coherence. This line of inquiry was further expanded upon by Resnick, Alfayez, Im, and Gilbert (2023), who investigated the reliability of crowdworkers in evaluating the accuracy of online articles, particularly in the context of political content.

3.4. Polarization and ideological analysis

Political polarization has become increasingly prominent in recent years, reshaping the dynamics of political discourse and ideological divisions. Akoglu (2014) presented a novel approach to quantify political polarity within opinion datasets. The proposed algorithm, "signed polarity propagation (SPP)", harnesses signed bipartite networks to classify individuals into political camps liberal or conservative, and rank both individuals and subjects based on the magnitude of their polarity. This innovative method showcased its effectiveness on real political datasets, offering a fresh perspective on the quantification of political polarization. Shi, Mast, Weber, Kellum, and Macy (2017) take a unique approach by analyzing cultural fault lines in the United States using Twitter data. Through an examination of co-following patterns on Twitter, the study measures the extent to which political divisions manifest in social media. It explores alignment and polarization across various cultural domains beyond traditional political indicators, providing insights into the multifaceted nature of polarization in contemporary society. As the digital landscape continued to evolve, Heatherly, Lu, and Lee (2017) navigated the intricate relationship between social networking sites (SNSs) and political discussions among U.S. Democrats and Republicans. Their research delved into the realms of cross-cutting versus like-minded discussions, further dissecting the roles played by affective polarization and party identification in shaping the contours of online political discourse.

In the study by Luttig (2017), the focus shifts towards exploring the relationship between authoritarianism and affective polarization in American politics. This research challenges the conventional belief that affective polarization primarily arises from psychological differences between Democrats and Republicans. Instead, it argues that authoritarianism is positively correlated with partisan extremism among both groups, suggesting shared psychological traits among strong partisans. Similarly, Johnston (2018) delves into the realm of affective polarization between Democrats and Republicans in the U.S., with a specific focus on economic ideology and authoritarianism. The study posits that economic preferences are influenced by the same personality divide that shapes preferences on other issues, ultimately leading to intense emotions in economic debates. Wojcieszak, Winter, and Yu (2020) examine the impact of social norms promoting open-mindedness on the selection of political content and its effect on affective polarization. Through two online experiments with American partisans, the study investigates how emphasizing norms of open-mindedness influences exposure to balanced and counter-attitudinal political content. The research seeks to determine whether promoting open-mindedness can alleviate the adverse effects of selective exposure, thus fostering informed citizenship in a polarized media landscape. In this era of unprecedented polarization, Manickam, Lan, Dasarathy, and Baraniuk (2019) proposed IdeoTrace—a pioneering framework designed to estimate the ideology of social media users and news websites. This framework was put to the test during the 2016 U.S. presidential election, utilizing matrix factorization techniques and social connections between users to trace shifts in user ideology over time. Their discoveries showed a growing divide between liberals and conservatives, exposing significant shifts happening in online political discussions.

One such political discussion explored by Demszky et al. (2019) was studying examines four distinct linguistic dimensions—topic choice, framing, affect, and illocutionary force—employing established lexical methods. Through the application of these techniques, the research uncovers evidence of high political polarization, driven by partisan differences in framing, particularly in tweets related to mass shootings. Ho, Kao, Li, Lai, and Chiu-Huang (2020) complement this by delving into the polarization of political opinions expressed by news media on Twitter. Their research focuses on the computational identification of news media's political opinions, based on the language used in their tweets. By exploring language disparities between left-wing and right-wing media, the study contributes to our understanding of how media outlets contribute to the polarization of political discourse. Additionally, Böttcher and Gersbach (2020) embark on an exploration of the mechanisms underlying political polarization in the U.S. Through mathematical frameworks and Bayesian Markov chain Monte Carlo techniques, the study meticulously analyzes empirical data on political polarization collected over several years. It investigates the spread of political and cultural ideas within populations leaning towards Democrats or Republicans and delves into the influential role of specific actors in shaping political opinions.

Further, to study the factors that can affect polarization, Kaufman, Kaufman, and Diep (2022) presented a statistical physics model aimed at comprehending the dynamics of political polarization in the United States. Their research takes into account anticipatory scenarios and external events, exploring the potential for alleviating or exacerbating polarization among Democrats, Republicans, and Independents. This interdisciplinary approach offers valuable insights into the complex and evolving landscape of political polarization. In a complementary vein, Diaz-Garcia and López (2023) utilizes entropy analysis to differentiate between non-organic and organic political content on Twitter. Their study distinguishes between original content and coordinated information operations, shedding light on the mechanisms that influence the spread of political narratives which further supports the examination of the influence of Russian interference on the 2016 U.S. Presidential election via Twitter by Badawy, Ferrara, and Lerman (2018) where they focused on users who re-shared posts from Russian troll accounts identified by the U.S. Congress, the study analyzes over 43 million election-related tweets from approximately 5.7 million users. The authors utilize label propagation to determine users' political ideology based on shared news sources, categorizing them as liberal or conservative. They find that conservatives played a significant role in amplifying troll content and explore the involvement of social bots. Additionally, geospatial analysis identifies regions where Russian troll content was particularly effective. Meanwhile, Jin (2023) further explore the reliability of crowdworkers in assessing the accuracy of news-like articles circulating on the internet. Their research investigates whether partisanship and inexperience influence judgment and tests various strategies to mitigate partisanship, contributing to more impartial content evaluation. Lastly, Zhu et al. (2023) introduce a Coevolving Latent Space Network with Attractors (CLSNA) model to understand the dynamics of partisan polarization on social media. This model distinguishes between positive and negative partisanship, offering a more comprehensive view of the intricate dynamics of ideological polarization.

3.5. Detecting political biases and orientation

The study of political biases and orientations is essential for understanding media landscapes and public discourse. Early contributions include Maynard and Funk (2012), who addressed the challenges of opinion mining from microposts and introduced advanced NLP techniques to extract political opinions from tweets. This foundational work laid the groundwork for subsequent research by demonstrating how sophisticated methods could effectively identify political biases in social media content.

Building on these early advancements, Saez-Trumper, Castillo, and Lalmas (2013) explored biases in both traditional and social media, focusing on metrics for selection, coverage, and statement bias. Their study provided insights into how biases manifest across different media platforms and how social media can amplify these biases. Following this, Tran (2020) introduced an unsupervised framework for estimating presentation bias at the source level. This approach emphasized source-level bias detection, offering an alternative to article-level analyses and showcasing the potential of unsupervised methods to enhance bias detection.

More recent developments include Gordon, Babaeianjelodar, and Matthews (2020), who adapted word embedding techniques to quantify political bias in tweets, extending the understanding of biases beyond binary axes. D'Alonzo and Tegmark (2022) introduced an automated method to measure media bias by analyzing phrase frequencies, mapping newspapers into a two-dimensional bias landscape. This approach complements existing models by offering a structured method for bias detection. Additionally, Xiao et al. (2023) presented Polarity-aware Embedding Multi-task Learning (PEM), a model designed to identify political biases in Twitter entities and hashtags. Lastly, Ness, Fatima, and Oghaz (2023) focused on quantifying political bias within mainstream media outlets using NLP and ML, systematically evaluating media sources to reveal biases influencing news reporting and framing.

3.6. Ideological dynamics in digital political discourse

The exploration into how ideological differences shape online political discourse commences with a reexamination of classic political theories in the digital realm. Le, Boynton, Mejova, Shafiq, and Srinivasan (2017) provide a nuanced analysis of sentiment trends, candidate discussions, and the influence of party affiliation, personality, and policy during the 2016 U.S. presidential primaries. Concurrently, Sterling, Jost, and Hardin (2019) delve into the ideological disparities between liberals and conservatives regarding their perceptions of a good society, uncovering divergent values and priorities across the ideological spectrum. These studies underscore the significance of digital platforms as arenas for expressing and negotiating diverse viewpoints, laying the foundation for deeper exploration into the intricate relationship between political party attachment and ideological commitment.

Barber and Pope (2019)'s investigation sheds light on the complex interplay between party loyalty and ideological principles, revealing how individuals prioritize party allegiance over ideological convictions in response to cues from influential party leaders. Prakasam and Huxtable-Thomas (2021) delve into the affordances of Reddit as a digital platform for constructing political narratives and identities. Their examination of the *r/The_Donald* community uncovers the role of digital spaces in facilitating the dissemination of political narratives and the reinforcement of ideological beliefs, underscoring the significance of digital platforms in shaping contemporary political discourse.

Hashemi (2021)'s proposal of a data-driven framework for coding political tweeting, disinformation, and extremism on online social networks provides a comprehensive approach to understanding ideological dynamics in digital political discourse. By manually classifying tweets based on their intent, focus, and political affiliation, Hashemi identifies various classes of intent, offering insights into the diverse range of ideological narratives present in online spaces during significant political events. This research contributes significantly to our understanding of how ideological differences manifest in digital discourse and the implications for political engagement and information dissemination in the digital age.

4. Findings RQ2: Advancements in identifying ideological extremism in U.S. social networks

The escalating prevalence of political extremism and radicalization in the United States has become a pressing societal issue. Understanding and effectively identifying extremist ideologies within online spaces present intricate challenges. Nonetheless, addressing these challenges is paramount for fostering a healthy digital environment and safeguarding societal well-being. This exploration focuses on existing ML, NLP, DL, graph-based methods, dictionary-based methods, and statistical methodologies tailored for the detection and classification of political extremism within U.S. social networking sites. By organizing research findings into coherent themes, we endeavor to provide a comprehensive overview of the field's advancements.

4.1. Understanding extremist opinions and online discussions

To effectively combat and address the rise of political extremism in online discussions, gaining insights into the infiltration of extremist opinions is crucial. Yang and Chen (2012) introduced a pioneering partially supervised learning approach to identify radical opinions within hate group web forums. Recognizing the challenge of obtaining labeled data for machine learning models, their methodology circumvented this obstacle by employing a labeling heuristic, facilitating the extraction of high-quality examples of extremist content from unlabeled datasets. This innovative approach significantly contributed to the identification of radical ideologies within online forums.

Furthermore, to explore diverse opinions, Wang, Wang, Erlandsson, Wu, and Faris (2013) delved into the factors that influence user participation in online newsgroups and the impact of feedback with different opinions. While their primary focus was on user behavior in online communities, their research provided insights into the interplay between diverse opinions and user engagement with political content. Understanding the impact of diverse opinions is a critical aspect of identifying and classifying political extremism in online discussions. Wang et al.'s study underscored the intricate relationship between the expression of varying viewpoints and their effects on user behavior. Further, extending the exploration into the behavior and psychology of political extremists on social media, Alizadeh, Weber, Cioffi-Revilla, Fortunato, and Macy (2019) conducted a comprehensive analysis of Twitter data. Their study focused on 10,000 political extremists associated with the alt-right and Antifa, comparing them with 5000 liberal and 5000 conservative users. The aim was to investigate differences in emotional expression and moral foundations among these groups. This extensive research provided valuable insights into the language and behavior of political extremists on Twitter, contributing to our understanding of extremism's psychological underpinnings.

Building upon these insights, Kong, Booth, Bailo, Johns, and Rizoiu (2022) adopted a mixed-method strategy to address the multifaceted challenge of understanding the manifestation and propagation of extremist opinions in online discussions. Integrating qualitative research, data collection from social media platforms, and advanced machine learning techniques, their approach aimed to uncover the emergence and co-occurrence of extreme opinions within online discourse. Through this comprehensive methodology, Kong et al. sought to gain deeper insights into the dynamics of extremist discourse and its interaction within online communities.

4.2. Extremist content analysis and detection

The proliferation of extremist content on the internet, particularly within the United States, has ignited concerns about its societal impact and prompted the need for effective detection and mitigation strategies. Wong, Frank, and Allsup (2015) conducted a content analysis of white supremacist online forums, unveiling these platforms' roles in information dissemination, recruitment, and networking among white supremacists. This exploration illuminated the offline implications of online hate speech and propaganda, underscoring the necessity of understanding extremist online behaviors. Addressing the identification of extremism in social media, Bhattacharjee, Balantrapu, Tolone, and Talukder (2017) introduced a dynamic learning framework designed to detect extremist content efficiently. Their framework incorporated context information and optimization methods, presenting a robust approach to identify extremist or criminal content amidst the vast volume of social media posts. In addition, Rudinac, Gornishka, and Worring (2017) proposed a multimodal approach to categorize user posts, focusing on violent online political extremism content. Using graph convolutional networks, they integrate text, visual content, and user interactions for analysis at a high semantic level. By applying entity linking and semantic concept detection, the study aims to classify extremist posts from Stormfront, demonstrating the potential of graph convolutional networks for multimedia classification and aiding qualitative data analysis of extremist content. Further, Owoeye and Weir (2018) presented a study detailing the development of an automated system called the Composite Method for classifying extremist web pages. This system integrates semantic and syntactic features of web page content, leveraging sentiment analysis and textual analysis tools. The findings indicate the superiority of the Composite Method over previous sentiment rule-based methods in terms of accuracy and efficiency, highlighting its potential as a valuable tool in combating online extremism. Similarly, Ribeiro, Ottoni, West, Almeida, and Meira (2020) scrutinized radicalization pathways on YouTube, revealing evidence of user radicalization facilitated by the platform's recommendation algorithms. Their findings emphasized the urgent need for proactive measures to counter the dissemination of extremist narratives online.

Consequently, Ai et al. (2021) investigated the popularity and persuasiveness of group videos with right- and left-leaning ideologies across various online platforms (YouTube, Bitchute, 4Chan, and Vimeo). Their study provided insights into the features and content that drive the spread of extremist ideologies, informing strategies to mitigate their influence. Moreover, Fahim and Gokhale (2021) focused on identifying social media content supporting the Proud Boys, employing machine learning models to distinguish extremist content and enhance understanding of right-wing extremism online.

Expanding the scope of analysis, Wang et al. (2021) conducted a multi-platform (Twitter, Reddit, 4chan, and Gab) assessment of political news discussions across diverse web communities, the trustworthiness of shared news stories, shedding light on the influence of different platforms in shaping political discourse. Similarly, Sipka, Hannak, and Urman (2022) compared QAnon-related content (volume, language usage, and context) across multiple social media platforms (Parler, Gab, and Twitter), offering valuable insights into the presence and nature of QAnon-related extremism online. Complementing these studies, Ebner, Kavanagh, and Whitehouse (2022) evaluated the national security threat posed by the QAnon movement through a combination of quantitative and qualitative analysis of online communication channels, identifying linguistic markers associated with violence risk.

Recently, Gaikwad et al. (2023) emphasized the importance of monitoring and countering extremism on social media by introducing a balanced multi-ideology extremism text dataset. Leveraging deep learning techniques, their research facilitated the detection and classification of extremist content across various ideological spectrums such as propaganda and radicalization, highlighting the significance of advanced machine learning methods in combating online extremism. Likewise, Ravi, Vela, Jenaway, and Windisch (2023) presented a novel approach to measuring threats in social media comments, particularly targeting voting and public officials in the US. By proposing a comprehensive threat level scale and collecting a vast dataset of 1.3 million Telegram responses, the study explores the use of AI-human annotation systems for efficient threat detection. Findings show promising results, indicating the potential of the GPT-2 model in cost-effective threat monitoring. The research contributes to understanding online threats and suggests strategies for continuous threat-level monitoring and enhancement. Through these collective efforts, researchers strive to develop comprehensive strategies to detect, mitigate, and counter the proliferation of extremist content in online environments.

4.3. Hate speech detection and classification

In the digital age, the proliferation of hate speech and extremist content on social media platforms has raised significant concerns, prompting researchers to develop innovative methods for their detection and classification. Melton et al. (2020) tackled the challenge of hate speech detection across multiple platforms, introducing a deep learning framework that combined various models and leveraged transfer learning and weak supervision techniques. Their approach showcased the efficacy of advanced machine learning methods in combating online extremism, emphasizing tailored approaches to differentiate hate speech from other forms of online discourse.

Focusing on specific hate groups, Simons and Skillicorn (2020) focused on distinguishing extremist rhetoric from potential extremist violence within online content, particularly in white supremacist forums. Their predictive models for intent and abusive language detection contributed to identifying posts indicating a desire for violent action, highlighting the nuanced nature of extremist content classification. Alatawi, Alhothali, and Moria (2021) concentrated on detecting white supremacist hate speech on Twitter, achieving high accuracy in identifying hate speech within extremist groups through deep learning and natural language processing techniques, underscoring the necessity of specialized methods for specific types of extremism. Additionally, Arviv, Hanouna, and Tsur (2021) investigated the use of symbols on social media platforms by alt-right members, white supremacists, and trolls, with a focus on targeting individuals of Jewish heritage and associated antisemitism. Their study involved constructing a dataset to examine racist online communities, employing natural language processing and network analysis to explore aspects such as disambiguating hate speech, network structure, hate intersectionality, linguistic variations of symbols, and the participation of the Internet Research Agency (IRA).

Building upon these efforts to detect and classify harmful content, Ali et al. (2023) proposed a novel approach combining deep learning techniques with graph algorithms to detect hate content on platforms like Twitter. Their research not only classified hate speech but also identified the communities responsible for spreading such content, offering insights into the network dynamics of extremism online. Complementing these efforts, Agnes, Solomon, and Tamilmaran (2023) addressed the broader goal of maintaining a positive online environment by identifying offensive language and vulgarity. Leveraging a Bidirectional LSTM model, their research aimed to classify Twitter comments into offensive and non-offensive categories, recognizing the intertwined nature of abusive content with extremist rhetoric.

Furthermore, Apostolopoulos, Liakos, and Delis (2022) introduced a social-aware deep learning approach for hate speech detection, incorporating social features to enhance classification accuracy and acknowledging the role of social interactions in the propagation of hate speech. Ajala et al. (2022) provided a comprehensive analysis of far-right extremist content on Twitter, employing AI and content analysis techniques to offer insights into various types of extremism, levels of radicalization, sentiment analysis, and the identification of opinion leaders within the far-right extremist community, shedding light on the multifaceted nature of extremism on social media platforms.

4.4. Understanding radicalization and beliefs

Effectively combating political extremism within the United States necessitates a comprehensive understanding of the values, attitudes, and beliefs that underpin extremist movements. Agarwal et al. (2014) conducted a study exploring the values of political movements like the Tea Party and Occupy Wall Street in relation to their use of technology, highlighting the importance of understanding ideological underpinnings in analyzing extremist groups. Similarly, Bevensee and Ross (2018) focused on the Alt-Right movement, examining its involvement in violent extremism and disinformation campaigns through social media data mining. Their research sheds light on the ideological foundations and dissemination strategies employed by extremist movements, emphasizing the need for nuanced analysis of influences as well as motivating factors, including geopolitical strategies.

Additionally, Barfar (2019)'s study investigated cognitive and affective responses to political disinformation on Facebook, offering insights into individuals' reactions to false information. It provided relevant insights into understanding how misinformation impacts belief systems. Furthermore, Grover and Mark (2019)'s paper aimed to identify warning signs of ideological radicalization within the alt-right community on Reddit, offering valuable insights into potential indicators of extremism within online communities and the ideological evolution of extremist groups.

Understanding the spread and evolution of extremist ideologies is crucial, as highlighted by Qi et al. (2010)'s research, which delves into the internet network structure of extremist political movements' web pages, seeking insights into their development and connections. The study introduces the quasi-clique merger method, a hierarchical clustering algorithm, to assess similarities among extremist web pages using their bi-directional hyperlink structure. By organizing web pages into communities, this method enables a deeper understanding of their relationships and interactions. Similarly, Youngblood (2020)'s paper modeled the spread of far-right radicalization in the United States, employing an epidemiological approach to provide insights into the factors contributing to radicalization and emphasizing the role of online and physical organizing in recruitment. Additionally, Diab, Jagdagdorj, Ng, Lin, and Yoder (2023) examined the crossover of white supremacist propaganda between online and offline spaces, emphasizing the importance of tracking the movement of extremist ideologies across different environments.

Extremist ideologies transcend traditional political spectrums, as demonstrated by Jones (2023)'s exploration of the global reach of the QAnon conspiracy theory. By delving into far-right and millennialist elements of QAnon, the study expanded our understanding of extremist narratives, highlighting their adaptability and evolving nature beyond conventional boundaries. These diverse research efforts underscore the multifaceted nature of political extremism and the necessity of comprehensive approaches to understand and counteract its influence.

4.5. Social media, political polarization and extremism

In the contemporary digital landscape, the intricate interplay between social media, political polarization, and extremism has emerged as a prominent concern. Swann and Husted (2017) examined the evolution of Occupy Wall Street (OWS) from a physical protest movement to an online presence, illustrating how the transition to digital platforms altered the movement's organizational dynamics. By analyzing Facebook activity, they highlighted the shift from participatory to more centralized practices, suggesting that social media, particularly Facebook, may undermine certain organizational principles. This underscores the transformative role of online platforms in reshaping the dynamics of political movements, pertinent to understanding the digital manifestations of extremism.

Bryanov, Vasina, Pankova, and Pakholkov (2021) investigated the repercussions of deplatforming political figures like Donald Trump on alternative social media platforms such as Telegram. Their study explored the growth of right-wing communities on Telegram post-deplatforming, shedding light on the resilience of extremism in alternative digital spaces. Additionally, Matias, Costales, and Christian (2022) addressed the rising concern of cybercrime on social media, focusing on predicting cyberbullying and cyberthreats on Twitter. Their research underscored the imperative of identifying and mitigating cybercrimes, including extremist threats and harassment, facilitated through online platforms.

Moreover, Gaikwad, Ahirrao, Kotecha, and Abraham (2022) emphasized the importance of detecting and classifying extremism on social media platforms through a multi-ideology, multi-class approach. By developing a balanced extremism text dataset with multi-class labels, they aimed to accurately categorize extremist content into various forms using deep learning techniques. This approach acknowledges the diversity of extremist ideologies and underscores the significance of precise classification in combating extremism effectively.

Furthermore, Withers, Parrish, Terrell, and Ellis (2017) explored the relationship between the “Dark Triad” personality traits and deviant behavior on social networking sites (Facebook, Twitter, and Instagram), offering insights into the psychological underpinnings of online behavior and its implications for extremist content creation and dissemination. Additionally, Nguyen and Gokhale (2022) presented an efficient approach to identifying anti-government sentiment on Twitter during politically motivated protests, addressing the challenge of monitoring online sentiment amidst social unrest. This research underscores the importance of real-time monitoring and response to extremist sentiments expressed on social media during moments of political turmoil.

Exploring political polarization on social media platforms, Kovacs, Cotfas, and Delcea (2022) delved into unhealthy online discourse on Twitter, particularly in the aftermath of the January 6th events at the US Capitol. Employing machine learning, they classified tweets as healthy or unhealthy, shedding light on the role of online political discourse in influencing extremism. Concurrently, Rajendran et al. (2022) conducted a study focusing on the creation of an extremism dataset derived from tweets gathered during the U.S. Capitol riot. This dataset encompasses various forms of extremism, including propaganda, recruitment, radicalization, and non-extremism for detecting extremism on social media platforms.

In addition, understanding platform-specific dynamics is crucial, as highlighted by Lee and Pirim (2023)’s comparative study of content and user behavior on Twitter and Parler during the January 6, 2021 Capitol Riots. This work emphasized the need for tailored strategies to monitor and counter extremism across diverse social media platforms, recognizing the nuances in content generation and user engagement. Such comprehensive analyses provide essential insights into the intricate nexus between social media, political polarization, and extremism in the digital age.

5. Data extraction and synthesis

Data extraction is a fundamental aspect of any systematic literature review (SLR), facilitating the comprehensive analysis of research findings. In our review, we meticulously examined the 110 included studies, employing a standardized template for summarization. This process aided in comprehensively understanding the subject matter and identifying potential avenues for future research. The extracted data were organized and stored in spreadsheets, systematically categorizing information according to study title, year of publication, and summary. For each research question, Tables 3, 4, 5, 6, and 7 delineate various extraction categories, including social media platforms, techniques, datasets, tasks, and programming language/tool, contributing crucial insights into the landscape of predictive modeling and analysis.

5.1. Social media platforms

The summary provided in Table 3 offers insights into the primary social media platforms utilized in the reviewed studies. Notably, a significant proportion of studies relied on data gathered from Twitter, with Facebook emerging as the second most utilized platform, followed by survey data obtained both offline and online.

The prevalence of studies utilizing Twitter data highlights its significance as a rich source of information for social media analysis (Chen, Duan, & Yang, 2022). However, it is essential to consider Twitter data’s inherent biases and limitations, such as its demographic skew (Blank, 2017) and potential for echo chambers. While not as extensively used as Twitter, including Facebook data underscores the importance of considering multiple platforms to capture diverse perspectives and behaviors (Chen et al., 2022). However, accessing Facebook data comes with challenges due to privacy concerns (Egelman, Oates, & Krishnamurthi, 2011) and access restrictions. The incorporation of survey data, both online and offline, indicates a recognition of the need to supplement social media data with more traditional research methods (Chen et al., 2022). This approach allows researchers to validate findings and ensure a more comprehensive understanding of social phenomena.

5.2. Techniques

Table 4 offers a comprehensive overview of the techniques employed across the reviewed studies. Our analysis reveals a rich and varied landscape of methodologies, including ML, NLP, DL, graph-based methods, dictionary-based methods, and statistical approaches.

In our systematic review, we categorize neural network-based methods, such as convolutional neural networks and graph neural networks, under deep learning. This classification emphasizes deep learning as a specialized subset of machine learning that encompasses these advanced techniques. Meanwhile, classical algorithms that do not rely on neural networks, such as support vector machines and logistic regression, fall under machine learning. These methods are crucial for tasks involving the classification and prediction of biases and orientations.

Natural language processing is another vital category, including essential text-processing techniques like tokenization, stemming, and lemmatization. These methods are fundamental for the effective analysis and interpretation of textual data. We also cover various graph-based analytical approaches within the graph-based methods category, including similarity search methods. Dictionary-based methods are explored for their use of predefined dictionaries in detecting bias and performing sentiment analysis. Additionally, we address statistical techniques, which encompass a range of tools and methods for data analysis and hypothesis testing.

To ensure clarity, we have included algorithms in all relevant categories when they fit multiple classifications. For example, articles employing graph neural networks are categorized under graph-based and deep-learning methods; when these methods involve text processing, they are also included in the NLP section.

The diverse application of statistical methods across studies underscores their foundational role in analyzing social media data. Statistical techniques offer robust interpretability (Daoud & Dubhashi, 2023) and are frequently used for hypothesis testing, complementing more complex machine learning approaches. On the other hand, the adoption of deep learning techniques signifies a growing trend towards leveraging sophisticated neural network architectures for tasks like sentiment

Table 3
Social media platforms used in the studies.

Media platform	RQ1	RQ2
Facebook	Alashri et al. (2016), Chiu and Hsu (2018), González-Bailón et al. (2023), Morris et al. (2020), Ribeiro et al. (2018)	Agarwal et al. (2014), Barfar (2019), Bevensee and Ross (2018), Kong et al. (2022), Melton et al. (2020), Swann and Husted (2017), Wang et al. (2013), Withers et al. (2017)
Twitter	Badawy et al. (2018), Chen et al. (2021), Colleoni et al. (2014), Decter-Frain and Barash (2022), Demszky et al. (2019), Diaz-Garcia and López (2023), Fagni and Cresci (2022), Fichman and Akter (2023), Golbeck and Hansen (2011), Gordon et al. (2020), Hashemi (2021, 2023), Himelboim et al. (2013), Ho et al. (2020), Lahoti et al. (2018), Le, Boynton, et al. (2017), Le, Shafiq, and Srinivasan (2017), Manickam et al. (2019), Maynard and Funk (2012), Morgan et al. (2013), Morris et al. (2020), Olteanu et al. (2022), Pennacchiotti and Popescu (2011), Preoţiuc-Pietro et al. (2017), Ramaciotti Morales (2022), Ribeiro et al. (2018), Saez-Trumper et al. (2013), Shi et al. (2017), Sterling et al. (2019, 2019), Stier (2016), Tien et al. (2020), Wong et al. (2016), Xiao et al. (2023), Zhu et al. (2023)	Agarwal et al. (2014), Agnes et al. (2023), Ajala et al. (2022), Alatawi et al. (2021), Ali et al. (2023), Alizadeh et al. (2019), Apostolopoulos et al. (2022), Arviv et al. (2021), Bevensee and Ross (2018), Bhattacharjee et al. (2017), Fahim and Gokhale (2021), Gaikwad et al. (2022, 2023), Jones (2023), Kong et al. (2022), Kovacs et al. (2022), Lee and Pirim (2023), Matias et al. (2022), Melton et al. (2020), Nguyen and Gokhale (2022), Rajendran et al. (2022), Sipka et al. (2022), Wang et al. (2021), Withers et al. (2017)
4chan		Ai et al. (2021), Melton et al. (2020), Wang et al. (2021)
TikTok	Medina Serrano et al. (2020)	Bryanov et al. (2021)
Telegram	Decter-Frain and Barash (2022)	Ebner et al. (2022), Ravi et al. (2023)
Instagram		Bevensee and Ross (2018), Withers et al. (2017)
Reddit	Alkiek et al. (2022), Botzer and Weninger (2023), Decter-Frain and Barash (2022), Prakasam and Huxtable-Thomas (2021), Ravi et al. (2022), Zhu et al. (2023)	Grover and Mark (2019), Wang et al. (2021)
Parler	Xiao et al. (2023)	Lee and Pirim (2023), Sipka et al. (2022)
Discord		Ebner et al. (2022)
Gab		Bevensee and Ross (2018), Ebner et al. (2022), Melton et al. (2020), Sipka et al. (2022), Wang et al. (2021)
YouTube		Ai et al. (2021), Bevensee and Ross (2018), Kong et al. (2022), Ribeiro et al. (2020), Withers et al. (2017)
Bitchute		Ai et al. (2021)
Vimeo		Ai et al. (2021)
News and web pages	Alzhrani (2022), D'Alonzo and Tegmark (2022), Hossain et al. (2018), Liu et al. (2022), Ness et al. (2023), Resnick et al. (2023), Ribeiro et al. (2018), Saez-Trumper et al. (2013), Tran (2020), Zerrer and Engelmann (2022)	Agarwal et al. (2014), Alatawi et al. (2021), Bevensee and Ross (2018), Bhattacharjee et al. (2017), Diab et al. (2023), Owoeye and Weir (2018), Rudinac et al. (2017), Simons and Skillicorn (2020), Wong et al. (2015), Yang and Chen (2012)
Study	Adamic and Glance (2005), Akoglu (2014), Malouf and Mullen (2008), Shaw and Benkler (2012)	Qi et al. (2010, 2010), Youngblood (2020)
Survey (Online)	Barber and Pope (2019), Heatherly et al. (2017), Kaufman et al. (2022), Kaye and Johnson (2016), Martel et al. (2022), Morris et al. (2020), Noel (2016), Resnick et al. (2023), Wojcieszak et al. (2020)	
Survey (Offline)	Böttcher and Gersbach (2020), Jin (2023), Johnston (2018), Lee (2005), Luttig (2017), Mason and Wronski (2018), Neo (2021), Preoţiuc-Pietro et al. (2017), Simas (2018)	Withers et al. (2017)

analysis and user classification. Despite their computational demands, deep learning models excel in handling unstructured data, such as text and images (Zhang, Yang, Chen, & Li, 2018).

Machine learning techniques, with their capacity for automated decision-making, are particularly valuable for applications such as political leaning and extremism classification. By learning from labeled data (Zhang et al., 2018), these methods enhance the accuracy and efficiency of predictive tasks. Concurrently, NLP methods like sentiment analysis and named entity recognition provide critical insights from textual data. Additionally, graph-based approaches reveal the intricate interconnectedness of social media data, offering insights into community structures and information diffusion processes (Das & Biswas, 2021).

These classifications and insights offer a clearer and more comprehensive understanding of the methodologies employed in the studies reviewed.

Table 4
Techniques used in the studies.

Method	RQ1	RQ2
Deep Learning	Alkiek et al. (2022), Alzhrani (2022), Fagni and Cresci (2022), Hashemi (2023), Liu et al. (2022), Ness et al. (2023), Ravi et al. (2022), Tran (2020), Xiao et al. (2023)	Agnes et al. (2023), Alatawi et al. (2021), Ali et al. (2023), Apostolopoulos et al. (2022), Arviv et al. (2021), Fahim and Gokhale (2021), Gaikwad et al. (2022, 2023), Kong et al. (2022), Kovacs et al. (2022), Melton et al. (2020), Nguyen and Gokhale (2022), Rajendran et al. (2022), Ravi et al. (2023), Rudinac et al. (2017), Simons and Skillicorn (2020), Wang et al. (2021)
Machine Learning	Chiu and Hsu (2018), D'Alonzo and Tegmark (2022), Fagni and Cresci (2022), Malouf and Mullen (2008), Manickam et al. (2019), Morgan et al. (2013), Morris et al. (2020), Olteanu et al. (2022), Pennacchiotti and Popescu (2011), Ravi et al. (2022), Tien et al. (2020), Wong et al. (2016), Zerrer and Engelmann (2022)	Agnes et al. (2023), Ai et al. (2021), Bhattacharjee et al. (2017), Bryanov et al. (2021), Fahim and Gokhale (2021), Kong et al. (2022), Kovacs et al. (2022), Matias et al. (2022), Nguyen and Gokhale (2022), Owoeye and Weir (2018), Qi et al. (2010), Rajendran et al. (2022), Ravi et al. (2023), Rudinac et al. (2017), Sipka et al. (2022), Wang et al. (2021), Yang and Chen (2012)
Natural Language Processing	Alashri et al. (2016), Alkiek et al. (2022), Badawy et al. (2018), Botzer and Weninger (2023), Chen et al. (2021), Chiu and Hsu (2018), Colleoni et al. (2014), Demszky et al. (2019), Diaz-Garcia and López (2023), Gordon et al. (2020), Hashemi (2023), Hossain et al. (2018), Liu et al. (2022), Malouf and Mullen (2008), Maynard and Funk (2012), Medina Serrano et al. (2020), Ness et al. (2023), Olteanu et al. (2022), Pennacchiotti and Popescu (2011), Preoţiuc-Pietro et al. (2017), Ramaciotti Morales (2022), Ravi et al. (2022), Saez-Trumper et al. (2013), Sterling et al. (2019), Stier (2016), Xiao et al. (2023)	Agnes et al. (2023), Ajala et al. (2022), Alatawi et al. (2021), Ali et al. (2023), Apostolopoulos et al. (2022), Bevensee and Ross (2018), Ebner et al. (2022), Gaikwad et al. (2022, 2023), Grover and Mark (2019), Kovacs et al. (2022), Lee and Pirim (2023), Matias et al. (2022), Melton et al. (2020), Nguyen and Gokhale (2022), Owoeye and Weir (2018), Ravi et al. (2023), Rudinac et al. (2017), Simons and Skillicorn (2020), Yang and Chen (2012)
Graph	Akoglu (2014), Böttcher and Gersbach (2020), Botzer and Weninger (2023), Colleoni et al. (2014), Decter-Frain and Barash (2022), Himelboim et al. (2013), Lahoti et al. (2018), Manickam et al. (2019), Preoţiuc-Pietro et al. (2017), Ramaciotti Morales (2022), Tien et al. (2020), Wong et al. (2016)	Ali et al. (2023), Arviv et al. (2021), Bhattacharjee et al. (2017), Qi et al. (2010), Rudinac et al. (2017), Wang et al. (2013)
Dictionary	Fichman and Akter (2023), Ho et al. (2020), Preoţiuc-Pietro et al. (2017)	Alizadeh et al. (2019), Barfar (2019), Grover and Mark (2019)
Statistics	Adamic and Glance (2005), Alashri et al. (2016), Barber and Pope (2019), Chen et al. (2021), Diaz-Garcia and López (2023), Golbeck and Hansen (2011), González-Bailón et al. (2023), Hashemi (2021), Heatherly et al. (2017), Jin (2023), Johnston (2018), Kaufman et al. (2022), Kaye and Johnson (2016), Le, Boynton, et al. (2017), Le, Shafiq, and Srinivasan (2017), Lee (2005), Luttig (2017), Martel et al. (2022), Mason and Wronski (2018), Neo (2021), Noel (2016), Prakasam and Huxtable-Thomas (2021), Ravi et al. (2022), Resnick et al. (2023), Ribeiro et al. (2018), Saez-Trumper et al. (2013), Shaw and Benkler (2012), Shi et al. (2017), Simas (2018), Wojcieszak et al. (2020), Zerrer and Engelmann (2022), Zhu et al. (2023)	Agarwal et al. (2014), Alizadeh et al. (2019), Diab et al. (2023), Jones (2023), Ribeiro et al. (2020), Swann and Husted (2017), Withers et al. (2017), Wong et al. (2015), Youngblood (2020)

5.3. Datasets

The summary provided in Table 5 outlines the datasets utilized in the reviewed studies. The availability and nature of datasets emerged as a critical aspect, with a predominant reliance on Twitter-specific or custom datasets. This focus on Twitter-specific datasets suggests an emphasis on phenomena unique to the platform, such as hashtag trends, retweet dynamics, and user interactions (Chen et al., 2022). Consequently, there is a growing need for increased publication of social datasets under standardized open licenses to enhance accessibility and drive further advancements in the field.

While the creation and utilization of custom datasets offer a tailored approach to address specific research questions or mitigate data biases (Olteanu, Castillo, Diaz, & Kıcıman, 2019), they may limit the generalizability, transparency, and reproducibility of findings. Moreover, the process requires substantial resources for data collection and annotation. In contrast, standardized datasets facilitate comparisons across studies and foster collaboration within the research community.

5.4. Tasks

Table 6 offers insights into the outcomes of the reviewed studies, particularly focusing on tasks such as political ideology and extremism detection. The most prevalent task identified was classification, followed by topic modeling, with content analysis often requiring manual and statistical processes. The utilization of diverse NLP methods underscores the importance of enhancing interpretability and explainability in future research endeavors.

Table 5
Datasets used in the studies.

Dataset	RQ1	RQ2
Facebook*	Alashri et al. (2016), Chiu and Hsu (2018), González-Bailón et al. (2023), Morris et al. (2020), Ribeiro et al. (2018)	Agarwal et al. (2014), Barfar (2019), Kong et al. (2022), Swann and Husted (2017), Wang et al. (2013), Withers et al. (2017)
Telegram*	Decter-Frain and Barash (2022)	Bryanov et al. (2021), Fahim and Gokhale (2021), Ravi et al. (2023)
Twitter*	Badawy et al. (2018), Chen et al. (2021), Colleoni et al. (2014), Decter-Frain and Barash (2022), Demszky et al. (2019), Diaz-Garcia and López (2023), Fagni and Cresci (2022), Fichman and Akter (2023), Golbeck and Hansen (2011), Gordon et al. (2020), Hashemi (2021, 2023), Himelboim et al. (2013), Ho et al. (2020), Lahoti et al. (2018), Le, Boynton, et al. (2017), Le, Shafiq, and Srinivasan (2017), Manickam et al. (2019), Maynard and Funk (2012), Morgan et al. (2013), Morris et al. (2020), Olteanu et al. (2022), Pennacchiotti and Popescu (2011), Preoţiuc-Pietro et al. (2017), Ramaciotti Morales (2022), Ribeiro et al. (2018), Saez-Trumper et al. (2013), Shi et al. (2017), Sterling et al. (2019), Stier (2016), Tien et al. (2020), Wong et al. (2016), Xiao et al. (2023), Zhu et al. (2023)	Agarwal et al. (2014), Agnes et al. (2023), Ajala et al. (2022), Alatawi et al. (2021), Ali et al. (2023), Alizadeh et al. (2019), Apostolopoulos et al. (2022), Arviv et al. (2021), Bevensee and Ross (2018), Bhattacharjee et al. (2017), Gaikwad et al. (2023), Jones (2023), Kong et al. (2022), Kovacs et al. (2022), Lee and Pirim (2023), Matias et al. (2022), Melton et al. (2020), Nguyen and Gokhale (2022), Rajendran et al. (2022), Sipka et al. (2022), Wang et al. (2021), Withers et al. (2017)
4chan*		Ai et al. (2021), Wang et al. (2021)
TikTok*	Medina Serrano et al. (2020)	
Instagram*		Withers et al. (2017)
Discord*		Ebner et al. (2022)
Reddit*	Alkiek et al. (2022), Botzer and Weninger (2023), Decter-Frain and Barash (2022), Prakasam and Huxtable-Thomas (2021), Ravi et al. (2022), Zhu et al. (2023)	Grover and Mark (2019), Wang et al. (2021)
Parler*		Lee and Pirim (2023), Sipka et al. (2022)
Gab*		Melton et al. (2020), Sipka et al. (2022), Wang et al. (2021)
YouTube*		Ai et al. (2021), Kong et al. (2022), Ribeiro et al. (2020), Withers et al. (2017)
Bitchute*		Ai et al. (2021)
Vimeo*		Ai et al. (2021)
IRA dataset	Diaz-Garcia and López (2023)	Arviv et al. (2021)
GDELT		Wang et al. (2021)
Parler posts	Xiao et al. (2023)	
All-the-news dataset	Tran (2020)	
TIMME	Xiao et al. (2023)	
ELECTION2020	Xiao et al. (2023)	Kovacs et al. (2022)
PIRUS		Youngblood (2020)
Pew Research Survey	Böttcher and Gersbach (2020), Kaufman et al. (2022), Ribeiro et al. (2018)	
ANES	Johnston (2018), Lee (2005), Luttig (2017)	
VolunteerScience*	Jin (2023)	
AMT*	Martel et al. (2022), Resnick et al. (2023)	
News and web pages*	D'Alonzo and Tegmark (2022), Liu et al. (2022), Martel et al. (2022), Ness et al. (2023), Resnick et al. (2023), Saez-Trumper et al. (2013), Zerrer and Engelmann (2022)	Agarwal et al. (2014), Alatawi et al. (2021), Bhattacharjee et al. (2017), Diab et al. (2023), Gaikwad et al. (2022, 2023), Owoeye and Weir (2018), Rudinac et al. (2017), Simons and Skillicorn (2020), Wong et al. (2015), Yang and Chen (2012)
YouGov	Noel (2016)	

* Denotes custom dataset.

The predominance of classification tasks underscores the significance of categorizing social media content based on attributes like sentiment, political ideology, and extremism (Sokolova & Lapalme, 2009). Classification facilitates automated decision-making and targeted interventions in areas like content moderation and recommendation systems. Meanwhile, the focus on topic modeling indicates a keen interest in identifying prevalent themes and discussions within social media data. Techniques such as Latent Dirichlet Allocation (LDA) facilitate content organization and contribute to understanding user interests and trends. Moreover, combining automated methods with human judgment ensures a nuanced interpretation of textual data, thereby mitigating algorithmic biases (Balayn, Lofi, & Houben, 2021).

Table 6
Tasks conducted in the studies.

Tasks	RQ1	RQ2
Classification	Akoglu (2014), Alkiek et al. (2022), Alzhrani (2022), Colleoni et al. (2014), D'Alonzo and Tegmark (2022), Diaz-Garcia and López (2023), Fagni and Cresci (2022), González-Bailón et al. (2023), Hashemi (2021, 2023), Ho et al. (2020), Liu et al. (2022), Malouf and Mullen (2008), Morgan et al. (2013), Ness et al. (2023), Olteanu et al. (2022), Pennacchiotti and Popescu (2011), Preoţiuc-Pietro et al. (2017), Ramaciotti Morales (2022), Ravi et al. (2022), Tran (2020), Xiao et al. (2023), Zerrer and Engelmann (2022)	Agnes et al. (2023), Ai et al. (2021), Alatawi et al. (2021), Ali et al. (2023), Apostolopoulos et al. (2022), Arviv et al. (2021), Bhattacharjee et al. (2017), Fahim and Gokhale (2021), Gaikwad et al. (2022, 2023), Kong et al. (2022), Kovacs et al. (2022), Matias et al. (2022), Melton et al. (2020), Nguyen and Gokhale (2022), Owoeye and Weir (2018), Rajendran et al. (2022), Ravi et al. (2023), Rudinac et al. (2017), Simons and Skillicorn (2020), Wang et al. (2013), Yang and Chen (2012)
Clustering	Akoglu (2014), Badawy et al. (2018), Decter-Frain and Barash (2022), Demszky et al. (2019), Fagni and Cresci (2022), Manickam et al. (2019), Noel (2016), Saez-Trumper et al. (2013), Zerrer and Engelmann (2022)	Ajala et al. (2022), Arviv et al. (2021), Qi et al. (2010), Rudinac et al. (2017), Sipka et al. (2022)
Sentiment analysis	Alashri et al. (2016), Botzer and Weninger (2023), Diaz-Garcia and López (2023), Ho et al. (2020), Jin (2023), Kaufman et al. (2022), Le, Boynton, et al. (2017), Liu et al. (2022), Martel et al. (2022), Preoţiuc-Pietro et al. (2017), Xiao et al. (2023), Zhu et al. (2023)	Agarwal et al. (2014), Ajala et al. (2022), Grover and Mark (2019), Lee and Pirim (2023), Simons and Skillicorn (2020), Withers et al. (2017), Wong et al. (2015)
Named Entity Recognition	Hossain et al. (2018), Maynard and Funk (2012), Xiao et al. (2023)	Ai et al. (2021), Ajala et al. (2022), Rudinac et al. (2017), Wang et al. (2021)
Dimensionality reduction	Decter-Frain and Barash (2022), Fagni and Cresci (2022), Liu et al. (2022), Manickam et al. (2019), Tien et al. (2020)	Ajala et al. (2022)
Topic modeling	Alashri et al. (2016), Medina Serrano et al. (2020), Pennacchiotti and Popescu (2011), Sterling et al. (2019), Wong et al. (2016)	Alizadeh et al. (2019), Bryanov et al. (2021)
Content analysis	Botzer and Weninger (2023), Fichman and Akter (2023), Gordon et al. (2020), Prakasam and Huxtable-Thomas (2021), Resnick et al. (2023)	Agarwal et al. (2014), Alizadeh et al. (2019), Apostolopoulos et al. (2022), Bevensee and Ross (2018), Diab et al. (2023), Ebner et al. (2022), Jones (2023), Swann and Husted (2017), Withers et al. (2017), Wong et al. (2015)
Community Detection	Tien et al. (2020)	Ali et al. (2023), Wang et al. (2021)
Influence Estimation	Böttcher and Gersbach (2020)	Wang et al. (2021)
Contagion process	Himelboim et al. (2013), Morgan et al. (2013)	Ribeiro et al. (2020), Youngblood (2020)

5.5. Programming language and tool

Finally, Table 7 delineates the programming languages and software utilized in the reviewed studies, showcasing a diverse array of choices. The preference for open-source languages and software reflects a commitment to transparency, collaboration, and reproducibility in research (Spies, 2013). Standalone tools tailored to specific research domains can streamline workflows, reduce dependencies on external libraries, and promote adopting best practices across the field.

6. Discussion

6.1. RQ1: Ideological orientation detection

The review of primary research articles on detecting ideological orientation in the United States, particularly on social networking sites, offers a comprehensive understanding of the methods employed to discern differences between liberal and conservative, right and left, or Democrat and Republican ideologies. By analyzing various themes (Section 3) within the literature, we can discover the current state of the art in tackling this question RQ1.

Firstly, in the realm of political-ideological analysis on social media, researchers have utilized ML, NLP, DL, graph-based methods, dictionary-based methods, and statistical approaches to gauge political preferences on platforms like Twitter. For instance, Golbeck and Hansen (2011) utilized follower connections to estimate political preferences, highlighting the alignment between media outlets' leanings and the preferences of their audiences. Similarly, Himelboim et al. (2013) mapped Twitter networks on contentious political topics, revealing the tendency of users to stick to ideological bubbles. Colleoni et al. (2014) classified Twitter users based on their political content sharing, emphasizing the higher political homophily among Democrats. Lahoti et al. (2018) proposed machine learning to model the liberal-conservative ideology space, aiming to identify ideological leaning for users and media sources.

Advanced techniques, such as deep learning approaches for political ideology detection, have also been explored. Malouf and Mullen (2008) extended NLP techniques to informal online political discussions, while Fagni and Cresci (2022) introduced an unsupervised deep learning approach to predict the political leaning of social media users.

Moreover, researchers have investigated media bias and perception, exploring how perceptions of media bias influence political discourse. Lee (2005) delved into public and political perceptions of media bias, while Adamic and Glance (2005) differentiated between liberal and conservative blogs' linking patterns. Kaufman et al. (2022) presented a statistical physics model to comprehend the dynamics of political polarization, considering anticipatory scenarios and external events.

Lastly, ideological dynamics in digital political discourse have been studied, with research focusing on the interplay between party loyalty and ideological principles. Barber and Pope (2019) revealed how individuals prioritize party allegiance over ideological convictions, while Prakasam and Huxtable-Thomas (2021) examined Reddit as a platform for constructing political narratives and identities.

Table 7
Programming languages and software used in the studies.

Languages/software	RQ1	RQ2
Python	Alkiek et al. (2022), Badawy et al. (2018), D'Alonzo and Tegmark (2022), Demszky et al. (2019), Fagni and Cresci (2022), Hashemi (2023), Ho et al. (2020), Liu et al. (2022), Maynard and Funk (2012), Ness et al. (2023), Olteanu et al. (2022), Ravi et al. (2022), Tran (2020)	Alatawi et al. (2021), Bevensee and Ross (2018), Jones (2023), Kong et al. (2022), Kovacs et al. (2022), Lee and Pirim (2023), Matias et al. (2022), Melton et al. (2020), Nguyen and Gokhale (2022), Rajendran et al. (2022), Ravi et al. (2023), Simons and Skillicorn (2020)
R	Sterling et al. (2019), Zhu et al. (2023)	Ebner et al. (2022), Youngblood (2020)
MATLAB	Tien et al. (2020)	
scikit-learn	Olteanu et al. (2022), Ravi et al. (2022)	Bhattacharjee et al. (2017), Kong et al. (2022), Kovacs et al. (2022), Nguyen and Gokhale (2022), Ravi et al. (2023)
WEKA	Chiu and Hsu (2018)	Owoeye and Weir (2018)
HuggingFace	Ravi et al. (2022)	Arviv et al. (2021), Kong et al. (2022), Kovacs et al. (2022), Ravi et al. (2023)
spaCy		Fahim and Gokhale (2021)
Stanford CoreNLP	Alashri et al. (2016)	Wang et al. (2021)
NLTK	Ness et al. (2023), Ravi et al. (2022), Tien et al. (2020)	Fahim and Gokhale (2021), Lee and Pirim (2023), Ravi et al. (2023)
LIWC	Ho et al. (2020), Preoţiuc-Pietro et al. (2017)	Ai et al. (2021), Alizadeh et al. (2019), Barfar (2019), Grover and Mark (2019), Wang et al. (2013)
Amazon Mechanical Turk	Barber and Pope (2019), Le, Shafiq, and Srinivasan (2017), Morris et al. (2020), Wojcieszak et al. (2020)	
Meaning Cloud		Kong et al. (2022)
Factiva		Kong et al. (2022)
LexisNexis		Kong et al. (2022)
Perspective API		Sipka et al. (2022)
NewsGuard API		Wang et al. (2021)
HateSonar		Grover and Mark (2019)
Manual analysis	Resnick et al. (2023)	
SPSS	Morgan et al. (2013)	
NOMINATE	Fagni and Cresci (2022), Noel (2016)	

Overall, the review highlights a range of methodologies employed to detect ideological orientation, including ML, NLP, DL approaches, as well as Graph-based methods, Dictionary-based methods, and Statistical analysis of online discourse patterns. These studies contribute to a nuanced understanding of ideological differences in digital spaces and their implications for political engagement and discourse.

Overall, the review highlights various methodologies employed to detect ideological orientation, ranging from ML, NLP, and DL approaches to graph-based, dictionary-based, and statistical analysis of online discourse patterns. These studies contribute to a nuanced understanding of ideological differences in digital spaces and their implications for political engagement and discourse.

6.1.1. Practical applications

By synthesizing a wide range of methodologies and research findings, our review provides a foundation for developing effective strategies and tools for understanding ideological dynamics in digital political discourse.

One practical implication of the reviewed research is the potential for developing more accurate and efficient machine-learning models for detecting political ideology on social media (Chiu & Hsu, 2018). By leveraging advanced techniques such as deep learning and natural language processing, researchers can enhance the capabilities of existing detection algorithms, enabling more precise identification of ideological biases.

Additionally, the findings from our literature review underscore the importance of promoting media literacy and critical thinking skills among social media users (Morgan et al., 2013). By equipping individuals with the tools to discern credible information from misinformation and identify ideological biases in online content, we can empower users to make informed decisions about the content they consume and share (Mason & Wronski, 2018). Educational initiatives to enhance media literacy can play a crucial role in combating the spread of extremist ideologies and fostering a more responsible and engaged online community.

Furthermore, the practical applications of ideological detection extend to various domains, including the development of solutions for news or content recommendation platforms (e.g., AllSides.com and Ground News), social media moderation, and intelligence gathering. By integrating insights from ideological detection research into these applications, organizations can foster improvements in operational efficiency and decision-making, ultimately contributing to a more transparent and informed digital landscape.

6.2. RQ2: Ideological extremism detection

In Section 4, we extensively addressed the question of the state of machine learning and natural language processing techniques in identifying ideological extremism within social networking sites in the United States. We have organized findings into several key themes, providing a comprehensive overview of advancements in this field.

To begin, our review highlights methodologies for understanding extremist opinions and online discussions. For instance, Yang and Chen (2012)'s partially supervised learning approach enabled the identification of radical opinions within hate group web forums, addressing challenges in obtaining labeled data for machine learning models. Additionally, studies such as Alizadeh et al. (2019)'s analysis of Twitter data contributed valuable insights into the language and behavior of political extremists.

Transitioning to approaches for analyzing and detecting extremist content, Wong et al.'s content analysis of white supremacist online forums illuminated the offline implications of online hate speech and propaganda (Wong et al., 2015). Bhattacharjee et al. (2017)'s dynamic learning framework efficiently detected extremist content on social media platforms.

Furthermore, we address hate speech detection and classification. Melton et al. (2020)'s deep learning framework showcased the efficacy of advanced machine learning methods in combating online extremism. Similarly, Simons and Skillicorn (2020)'s predictive models for intent and abusive language detection contributed to distinguishing extremist rhetoric from potential extremist violence.

In understanding radicalization and beliefs, Agarwal et al. (2014)'s exploration of political movements highlighted the importance of understanding ideological underpinnings in analyzing extremist groups. Additionally, studies such as Qi et al. (2010)'s investigation of the internet network structure of extremist political movements' web pages provided insights into their development and connections.

Finally, we examine the interplay between social media, political polarization, and extremism. Swann and Husted (2017)'s examination of Occupy Wall Street's transition to digital platforms highlighted the transformative role of online platforms in reshaping the dynamics of political movements. Similarly, studies like Bryanov et al. (2021)'s investigation of the growth of right-wing communities on alternative social media platforms post-deplatforming shed light on the resilience of extremism in alternative digital spaces.

Overall, our review demonstrates the diverse approaches and methodologies employed in identifying ideological extremism within social networking sites in the United States, showcasing the significance of advanced machine learning and natural language processing techniques in addressing this pressing societal issue.

6.2.1. Practical implications

The advancements in identifying ideological extremism in U.S. social networks offer practical implications for addressing real-world challenges. By leveraging deep learning and natural language processing techniques, researchers can develop more accurate and efficient machine-learning models for detecting extremist content on social media platforms. Studies such as Kong et al. (2022), which integrate qualitative research methods with advanced machine learning techniques, facilitate a multifaceted understanding of extremist discourse and its propagation in online communities. This combination allows for deeper insights into the emergence and co-occurrence of extreme opinions within online discourse, thereby informing more effective strategies for combating extremism.

Insights gleaned from the analysis of extremist content and user behavior on social media platforms contribute to the development of practical tools and strategies for monitoring and moderating online spaces. For instance, research such as Bhattacharjee et al. (2017), Wong et al. (2015) provides valuable insights into the roles of online forums and social media platforms in disseminating extremist content. Understanding the patterns and mechanisms underlying the spread of extremist ideologies enables platforms to implement more effective content moderation policies and algorithms, thereby mitigating the proliferation of harmful content.

Furthermore, research efforts such as Ribeiro et al. (2020) shed light on the impact of recommendation algorithms on user radicalization, emphasizing the need for proactive measures to counter the dissemination of extremist narratives online. Platforms can develop algorithms that prioritize the promotion of credible and diverse content while minimizing the amplification of extremist viewpoints, creating a safer and more responsible online environment for users.

Educational initiatives to enhance media literacy and critical thinking skills among social media users play a crucial role in combating the spread of extremist ideologies. Insights from studies such as Alizadeh et al. (2019), Melton et al. (2020) can inform the development of educational programs that equip individuals with the tools to discern credible information from misinformation and identify ideological biases in online content. By promoting media literacy and critical thinking skills, users can make informed decisions about the content they consume and share, thereby mitigating the influence of extremist propaganda and fostering a more responsible and engaged online community that upholds the values of democracy, diversity, and free expression.

6.3. Large language models in ideology and extremism detection

Large Language Models (LLMs) have become vital tools in identifying and classifying ideologies within digital discourse. Recent research, such as that by Ravi et al. (2022), leveraged natural language processing (NLP) techniques to analyze social media texts and distinguish between conservative and liberal content. By employing advanced classifiers and focusing on the linguistic characteristics of various political ideologies, this study provided an in-depth understanding of how ideology is communicated in digital communities. This type of ideological classification offers crucial insights for identifying the presence of extremist narratives within larger political conversations.

Expanding on this, LLMs have also been employed in the detection of threats and extremist behaviors in online environments. Ravi et al. (2023) introduced a threat-level scale to measure hostility in social media comments targeting voting and public officials. Utilizing the GPT-2 model alongside human-AI annotation systems, the study revealed the model's capacity for efficient and cost-effective threat detection, suggesting that NLP models could be adapted to monitor extremist threats more broadly. These findings underscore the potential of LLMs to serve as scalable tools for continuously monitoring and mitigating the rise of extremist content across social media platforms.

Further developments in ideology detection through LLMs have been realized by categorizing complex political beliefs beyond binary classifications. In Ravi and Vela (2024b), researchers introduced a dataset encompassing a wider range of ideological perspectives, including liberal, conservative, and more radicalized groups, sourced from various subreddits. The study found that while advanced models like transformers showed potential, simpler models such as support vector machines (SVM) with TF-IDF features outperformed others in capturing the nuanced differences in political ideologies. By applying these models to a comprehensive dataset, researchers demonstrated the value of combining NLP techniques with robust datasets for a more precise identification of extremist ideologies and their underlying narratives.

Moreover, the integration of LLMs into multi-platform assessments of radicalization pathways has proven crucial in understanding the spread of extremist ideologies across diverse online spaces. For instance, Ribeiro et al. (2020) highlighted how recommendation algorithms on YouTube facilitate user radicalization, a finding that aligns with broader concerns about algorithmic influence in the dissemination of extremist narratives. Similarly, research by Sipka et al. (2022), Wang et al. (2021) explored the role of multiple platforms in shaping political discourse and the prevalence of extremist content. These studies demonstrate that LLMs, when deployed across multiple platforms, are essential for uncovering the linguistic markers of extremism (RQ2) and contributing to content moderation strategies aimed at reducing the influence of radical ideologies online (RQ1).

6.4. Data extraction

6.4.1. Media platform

Our review on RQ1 revealed that Twitter emerged as the most frequently utilized platform for analyzing ideological orientation, with a significant number of studies (represented by citations) focusing on this platform. Facebook also garnered considerable attention from researchers, followed by platforms (Ravi & Vela, 2024b) such as Reddit, TikTok, Telegram, and Parler. In investigating RQ2, we observed a wide range of social media platforms studied in the literature. While Twitter and Facebook remained prominent, platforms like 4chan, TikTok, Telegram, Parler, Gab, YouTube, and Instagram also garnered attention, unlike in RQ1. Additionally, studies analyzed ideological content on news websites and web pages, utilizing both online and offline surveys. This reflects the importance of examining broader online spaces beyond traditional social media platforms and highlights the breadth of research conducted to understand ideological trends across various online spaces.

6.4.2. Techniques

Our review of the techniques used in the literature to address RQ1 and RQ2 reveals a diverse array of methodologies. We observed that NLP was the most commonly utilized technique for analyzing ideological orientation and extremism, with a substantial number of studies employing this method. ML techniques were also widely used, followed by DL and graph-based approaches. In addressing RQ1, there was a notable increase in the use of statistical techniques, reflecting a multifaceted approach to analyzing ideological content. This variety of techniques underscores the complexity of studying ideological orientation and extremism on social media platforms and highlights the importance of using a range of analytical tools to achieve comprehensive insights.

6.4.3. Datasets

In our observation of the datasets utilized in the literature to address RQ1 and RQ2, we discerned a variety of sources employed for analyzing ideological orientation and extremism on social media platforms. Notably, Twitter emerged as the most frequently utilized dataset, with a significant number of studies focusing on this platform for both RQ1 and RQ2. Similarly, Facebook and Reddit, along with news websites and web pages, were also extensively studied for both research questions, although to a lesser extent compared to Twitter. Additionally, custom datasets (*indicated) were used in some studies, reflecting efforts to gather specific data relevant to the research objectives. This underscores the need for standard and public datasets (Ravi & Vela, 2024a; Ravi & Yuan, 2024) to facilitate comparability and reproducibility across studies such as medical image (Hoogenboom et al., 2021; Kumar, Ravi, Mulay, Ram, & Sivaprakasam, 2018; Ravi, Selvaraj, Mulay, Ram, & Sivaprakasam, 2018) and signal analysis (Kamalakkannan, Rajkumar, Raj, & Devi, 2014). Some standard datasets include the IRA dataset, GDELT, TIMME, ELECTION2020, PIRUS, Pew Research Survey, ANES, VolunteerScience, AMT, and YouGov.

6.4.4. Tasks

In our observation of the techniques used in the literature to address RQ1 and RQ2, we discerned a variety of methodologies employed across different tasks. Classification emerged as the most commonly employed technique for both RQ1 and RQ2, with studies utilizing it to categorize and classify ideological content on social media platforms. Additionally, sentiment analysis and clustering techniques were frequently employed. Named Entity Recognition (NER), dimensionality reduction techniques, topic modeling, content analysis, and community detection were also utilized to varying degrees. This highlights the diverse range of analytical tools employed to study ideological orientation and social media use.

6.4.5. Tools

In our observation of the techniques used in the literature to address RQ1 and RQ2, we found that Python was the predominant language/software utilized for both tasks. Python was widely used for data processing, analysis, and machine learning implementations. R was also employed in some studies, albeit to a lesser extent compared to Python. Additionally, scikit-learn, NLTK, HuggingFace, and Stanford CoreNLP were among the commonly utilized libraries and tools within the Python ecosystem for natural language processing tasks related to ideological orientation analysis. Furthermore, Amazon Mechanical Turk was frequently used for collecting labeled data and manual analysis, while other software and APIs such as WEKA, LIWC, SPSS, and NOMINATE were utilized to a lesser extent.

6.5. Recommendations

6.5.1. Complexity of ideological analysis

The analysis of ideological orientation and extremism on social networking platforms in the United States has traditionally been constrained by a binary perspective, distinguishing primarily between liberal and conservative viewpoints. This dualistic approach significantly oversimplifies the complex landscape of political beliefs, which encompasses moderates, libertarians, and other nuanced ideologies that do not align perfectly with the conventional liberal-conservative spectrum. The reliance on binary classification neglects the diversity of political expressions and the subtle nuances that characterize individual ideological positions, including the use of specific phrases, symbols, or references that may not be explicitly political but signify a particular ideological leaning or sentiment.

Moreover, the focus on detecting extremism has predominantly targeted the identification of clear instances of hate speech, abusive language, radicalization, and white supremacy. While addressing these manifestations of extremism is undeniably essential, this narrow scope can overlook the more subtle and nuanced expressions of ideological beliefs and extremism. For instance, nuanced forms of extremism might not be expressed through overtly aggressive language but rather through coded messages, dog whistles, or the dissemination of certain conspiracy theories. These subtler expressions require a more refined approach to analysis, one that can interpret the varying meanings of terms or symbols across different contexts. This necessitates a move beyond simple keyword or sentiment analysis towards a more sophisticated and contextual understanding of ideological expressions, recognizing the complexity and multiplicity of political orientations beyond the binary framework.

6.5.2. Platform diversity

The research on detecting ideological orientation and extremism has predominantly focused on mainstream social networking sites like Twitter and Facebook, revealing significant insights into online ideological discourse. However, this concentration on a select few platforms presents a limitation in understanding the full scope of how ideologies are expressed and evolve across the digital landscape. Emerging platforms such as TikTok, Telegram, and Parler are beginning to receive scholarly attention, signaling a recognition of their growing importance. Yet, a comprehensive cross-platform analysis that encompasses these newer and less conventional platforms (depending on the platform's features, user base, and content moderation policies) remains a notable gap in the literature.

6.5.3. Dataset availability and standardization

A significant challenge within the field of detecting ideological orientation and extremism on social networking sites is the heavy reliance on specific datasets, notably from platforms like Twitter. While Twitter provides a valuable source of data due to its public API and widespread use for political discourse, this focus limits the diversity of datasets and potentially biases our understanding of online ideological landscapes. Furthermore, the use of custom datasets, though tailored to specific research questions, complicates the comparability and reproducibility of studies. There is a clear need for more standardized and publicly available datasets that span a variety of platforms and contexts to enhance the rigor and breadth of research in this area.

6.5.4. Longitudinal and dynamic analysis

The prevalent focus on static analytical approaches such as classification, sentiment analysis, and clustering in the study of ideological orientation on social media platforms points to a significant gap in understanding the dynamic nature of political ideologies. The evolution of ideological orientation over time, influenced by political events, shifts in public opinion, and changes in platform algorithms, remains underexplored. Addressing this gap through longitudinal and dynamic analysis can offer valuable insights into how ideologies shift, grow, or polarize within digital spaces, providing a deeper understanding of the fluidity and complexity of online political discourse.

6.5.5. Mitigating bias and enhancing fairness

In the pursuit of detecting ideological orientation on social media, the importance of mitigating bias and enhancing fairness cannot be overstated. Biases in dataset collection, model training, and analytical processes can significantly skew research outcomes, leading to inaccurate representations of ideological spectrums and potentially reinforcing existing stereotypes and inequalities. Addressing these biases requires a multifaceted approach that encompasses the development of methodologies designed to ensure the fairness and accuracy of ideological orientation detection. Moreover, the exploration of ideological orientation through multimodal data (e.g., text, images, videos) and the investigation of misinformation's impact are critical areas for developing a more nuanced understanding of online political discourse.

7. Conclusion

In this systematic literature review, we aimed to address the persistent challenges in understanding and combating online ideological extremism, particularly within the context of social networking sites in the United States. Despite significant advancements, three major research gaps have remained: the need for a comprehensive synthesis of existing methodologies (Research Gap 1), the necessity of temporal analysis and practical application of these methodologies (Research Gap 2), and the identification of ongoing challenges, unexplored areas, and recommendations for future research (Research Gap 3).

To bridge these gaps, we posed two research questions focused on synthesizing existing literature on techniques for detecting ideological orientation (RQ1) and extremism (RQ2). Our review of 110 primary research articles from 2005 to 2023 has provided a thorough thematic analysis, addressing the need for synthesis and temporal insight.

Our findings indicate that while substantial progress has been made, particularly in using NLP, ML, and other advanced methodologies, challenges such as platform diversity, dataset standardization, and mitigating bias remain critical areas for future work. Additionally, we identified a lack of longitudinal studies and dynamic analysis, essential for understanding extremist ideologies' evolution over time.

By systematically analyzing the existing literature, our review not only highlights the current state of research but also provides actionable recommendations for researchers and policymakers. These include the advancement of methodologies, fostering interdisciplinary collaborations, and exploring emerging platforms like TikTok and Telegram.

Moving forward, our work underscores the importance of continued innovation in the methodologies used to detect and mitigate ideological extremism on social media. We advocate for a focus on improving bias mitigation, enhancing fairness and transparency, and standardizing data practices to better equip researchers and policymakers in combating online extremism.

CRediT authorship contribution statement

Kamalakkannan Ravi: Conceptualization, Investigation, Methodology, Formal analysis, Visualization, Data curation, Writing – original draft. **Jiann-Shiun Yuan:** Supervision, Methodology, Formal analysis, Writing – review.

Open source statement

The research articles collection used in this review article is publicly available in a spreadsheet.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Azwad Tamir and Fatemeh Dashti Ahangar for their help in proofreading this paper. Article processing charges were provided in part by the UCF College of Graduate Studies Open Access Publishing Fund.

Data availability

No data was used for the research described in the article.

References

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on link discovery* (pp. 36–43).

Agarwal, S. D., Barthel, M. L., Rost, C., Borning, A., Bennett, W. L., & Johnson, C. N. (2014). Grassroots organizing in the digital age: Considering values and technology in tea party and occupy wall street. *Information, Communication & Society*, *17*(3), 326–341.

Agnes, S. A., Solomon, A. A., & Tamilmaran, D. J. C. (2023). Abusive comment detection in social media with bidirectional LSTM model. In *2023 5th international conference on smart systems and inventive technology* (pp. 1368–1373). IEEE.

Ai, L., Kathuria, A., Panda, S., Sahai, A., Yu, Y., Levitan, S. I., et al. (2021). Identifying the popularity and persuasiveness of right-and left-leaning group videos on social media. In *2021 IEEE international conference on big data* (pp. 2454–2460). IEEE.

Ajala, I., Feroze, S., El Barachi, M., Oroumchian, F., Mathew, S., Yasin, R., et al. (2022). Combining artificial intelligence and expert content analysis to explore radical views on Twitter: Case study on far-right discourse. *Journal of Cleaner Production*, *362*, Article 132263.

Akoglu, L. (2014). Quantifying political polarity based on bipartite opinion networks. In *Proceedings of the international AAAI conference on web and social media*: *vol. 8*, (no. 1), (pp. 2–11).

Alashri, S., Kandala, S. S., Bajaj, V., Ravi, R., Smith, K. L., & Desouza, K. C. (2016). An analysis of sentiments on Facebook during the 2016 US presidential election. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 795–802). IEEE.

Alatawi, H. S., Alhothali, A. M., & Moria, K. M. (2021). Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. *IEEE Access*, *9*, 106363–106374.

Ali, M., Hassan, M., Kifayat, K., Kim, J. Y., Hakak, S., & Khan, M. K. (2023). Social media content classification and community detection using deep learning and graph analytics. *Technological Forecasting and Social Change*, *188*, Article 122252.

Alizadeh, M., Weber, I., Cioffi-Revilla, C., Fortunato, S., & Macy, M. (2019). Psychology and morality of political extremists: evidence from Twitter language analysis of alt-right and Antifa. *EPJ Data Science*, *8*(1), 1–35.

Alkiek, K., Zhang, B., & Jurgens, D. (2022). Classification without (proper) representation: Political heterogeneity in social media and its implications for classification and behavioral analysis. In *Findings of the association for computational linguistics: ACL 2022* (pp. 504–522).

Alzhrani, K. M. (2022). Politicians-based deep learning models for detecting news, authors and media political ideology. *International Journal of Advanced Computer Science and Applications*, *13*(2).

Apostolopoulos, G. C., Liakos, P., & Delis, A. (2022). A social-aware deep learning approach for hate-speech detection. In *Asia-Pacific web (aPWeb) and web-age information management (WAIM) joint international conference on web and big data* (pp. 536–544). Springer.

Arviv, E., Hanouna, S., & Tsur, O. (2021). It'sa thin line between love and hate: Using the echo in modeling dynamics of racist online communities. In *Proceedings of the international AAAI conference on web and social media*: *vol. 15*, (pp. 61–70).

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 258–265). IEEE.

Balayn, A., Lofi, C., & Houben, G.-J. (2021). Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, *30*(5), 739–768.

Barber, M., & Pope, J. C. (2019). Does party trump ideology? Disentangling party and ideology in America. *American Political Science Review*, *113*(1), 38–54.

Barfar, A. (2019). Cognitive and affective responses to political disinformation in Facebook. *Computers in Human Behavior*, *101*, 173–179.

Behzadan, V., Aguirre, C., Bose, A., & Hsu, W. (2018). Corpus and deep learning classifier for collection of cyber threat indicators in Twitter stream. In *2018 IEEE international conference on big data* (pp. 5002–5007). IEEE.

Bevensee, E., & Ross, A. R. (2018). The alt-right and global information warfare. In *2018 IEEE international conference on big data* (pp. 4393–4402). IEEE.

Bhattacharjee, S. D., Balantrapu, B. V., Tolone, W., & Talukder, A. (2017). Identifying extremism in social media with multi-view context-aware subset optimization. In *2017 IEEE international conference on big data* (pp. 3638–3647). IEEE.

Blank, G. (2017). The digital divide among Twitter users and its implications for social research. *Social Science Computer Review*, *35*(6), 679–697.

Böttcher, L., & Gersbach, H. (2020). The great divide: drivers of polarization in the US public. *EPJ Data Science*, *9*(1), 32.

Botzer, N., & Weninger, T. (2023). Entity graphs for exploring online discourse. *Knowledge and Information Systems*, 1–19.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, *80*(4), 571–583.

Bryanov, K., Vasina, D., Pankova, Y., & Pakholkov, V. (2021). The other side of deplatforming: Right-wing Telegram in the wake of trump's Twitter ouster. In *International conference on digital transformation and global society* (pp. 417–428). Springer.

Chen, K., Duan, Z., & Yang, S. (2022). Twitter as research data: Tools, costs, skill sets, and lessons learned. *Politics and the Life Sciences*, *41*(1), 114–130.

Chen, W., Pacheco, D., Yang, K.-C., & Menczer, F. (2021). Neutral bots probe political bias on social media. *Nature Communications*, *12*(1), 5580.

Chiu, S.-I., & Hsu, K.-W. (2018). Predicting political tendency of posts on Facebook. In *Proceedings of the 2018 7th international conference on software and computer applications* (pp. 110–114).

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, *64*(2), 317–332.

Cortis, K., & Davis, B. (2021). Over a decade of social opinion mining: a systematic review. *Artificial Intelligence Review*, *54*(7), 4873–4965.

D'Alonzo, S., & Tegmark, M. (2022). Machine-learning media bias. *PLoS One*, *17*(8), Article e0271947.

Daoud, A., & Dubhashi, D. (2023). Statistical modeling: The three cultures. *Harvard Data Science Review*, *5*(1), https://hdsr.mitpress.mit.edu/pub/uo4hjcx6.

Das, S., & Biswas, A. (2021). Deployment of information diffusion for community detection in online social networks: a comprehensive review. *IEEE Transactions on Computational Social Systems*, *8*(5), 1083–1107.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*: *vol. 11*, (no. 1), (pp. 512–515).

Davis, N. T. (2017). *Partisanship, ideology, and the sorting of the American mass public*. Louisiana State University and Agricultural & Mechanical College.

Decter-Frain, A., & Barash, V. (2022). Using knowledge graphs to detect partisanship in online political discourse. In *International conference on complex networks and their applications* (pp. 50–61). Springer.

Demszky, D., Garg, N., Voigt, R., Zou, J., Shapiro, J., Gentzkow, M., et al. (2019). *Analyzing polarization in social media: Method and application to tweets on 21 mass shootings* (pp. 2970–3005). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1304.

Denyer, D., & Tranfield, D. (2009). Producing a systematic review.

Department of Justice (2018). Grand jury indicts thirteen Russian individuals and three Russian companies for scheme to interfere in the United States political system.

Diab, A., Jagdagdorj, B.-E., Ng, L. H. X., Lin, Y.-R., & Yoder, M. M. (2023). Online to offline crossover of white supremacist propaganda. In *Companion proceedings of the ACM web conference 2023* (pp. 1308–1316).

Diaz-Garcia, J. A., & López, J. A. D. (2023). All trolls have one mission: An entropy analysis of political misinformation spreaders. In *International conference on flexible query answering systems* (pp. 159–167). Springer.

Ebner, J., Kavanagh, C., & Whitehouse, H. (2022). The QAnon security threat. *Perspectives on Terrorism*, *16*(6), 62–86.

Egelman, S., Oates, A., & Krishnamurthi, S. (2011). Oops, I did it again: Mitigating repeated access control errors on facebook. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2295–2304).

Erbschloe, M. (2018). *Extremist propaganda in social media: A threat to homeland security*. CRC Press.

Fagni, T., & Cresci, S. (2022). Fine-grained prediction of political leaning on social media with unsupervised deep learning. *Journal of Artificial Intelligence Research*, *73*, 633–672.

Fahim, M., & Gokhale, S. S. (2021). Identifying social media content supporting proud boys. In *2021 IEEE international conference on big data* (pp. 2487–2495). IEEE.

Fernandez, M., Asif, M., & Alani, H. (2018). Understanding the roots of radicalisation on Twitter. In *Proceedings of the 10th ACM conference on web science* (pp. 1–10).

Fichman, P., & Akter, S. (2023). Trolling asymmetry toward Republicans and democrats and the shift from foreign to domestic trolling. *Telematics and Informatics*, Article 101998.

Gaikwad, M., Ahirrao, S., Kotecha, K., & Abraham, A. (2022). Multi-ideology multi-class extremism classification using deep learning techniques. *IEEE Access*, *10*, 104829–104843.

Gaikwad, M., Ahirrao, S., Phansalkar, S., Kotecha, K., Rani, S., et al. (2023). Multi-ideology, multiclass online extremism dataset, and its evaluation using machine learning. *Computational Intelligence and Neuroscience*, *2023*.

Gillies, J., & Cailliau, R. (2000). *How the web was born: The story of the World Wide Web*. USA: Oxford University Press.

Golbeck, J., & Hansen, D. (2011). Computing political preference among Twitter followers. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1105–1108).

González-Bailón, S., Lazer, D., Barberá, P., Zhang, M., Allcott, H., Brown, T., et al. (2023). Asymmetric ideological segregation in exposure to political news on Facebook. *Science*, *381*(6656), 392–398.

Gordon, J., Babaeianjelodar, M., & Matthews, J. (2020). Studying political bias via word embeddings. In *Companion proceedings of the web conference 2020* (pp. 760–764).

Grover, T., & Mark, G. (2019). Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In *Proceedings of the international AAAI conference on web and social media*: *vol. 13*, (pp. 193–204).

Hashemi, M. (2021). A data-driven framework for coding the intent and extent of political tweeting, disinformation, and extremism. *Information*, *12*(4), 148.

Hashemi, M. (2023). Geographical visualization of tweets, misinformation, and extremism during the USA 2020 presidential election using LSTM, NLP, and GIS. *Journal of Big Data*, *10*(1), 125.

Heatherly, K. A., Lu, Y., & Lee, J. K. (2017). Filtering out the other side? Cross-cutting and like-minded discussions on social networking sites. *New Media & Society*, *19*(8), 1271–1289.

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, *18*(2), 154–174.

Ho, S. M., Kao, D., Li, W., Lai, C.-J., & Chiu-Huang, M.-J. (2020). "On the left side, there's nothing right. On the right side, there's nothing left:" polarization of political opinion by news media. In *Sustainable digital communities: 15th international conference, iConference 2020, Boras, Sweden, March 23–26, 2020, proceedings 15* (pp. 209–219). Springer.

Holt, T. J., Freilich, J. D., & Chermak, S. M. (2022). Examining the online expression of ideology among far-right extremist forum users. *Terrorism and Political Violence*, *34*(2), 364–384.

Hoogenboom, S. A., Ravi, K., Engels, M. M., Irmakci, I., Keles, E., Bolan, C. W., et al. (2021). Missed diagnosis of pancreatic ductal adenocarcinoma detection using deep convolutional neural network. *Gastroenterology*, *160*(6, Supplement), S–18. http://dx.doi.org/10.1016/S0016-5085(21)00794-0, URL https://www.sciencedirect.com/science/article/pii/S0016508521007940.

Hossain, N., Tran, T. T. T., & Kautz, H. (2018). Discovering political slang in readers' comments. In *Proceedings of the international AAAI conference on web and social media*: *vol. 12*, (no. 1).

Jin, X. (2023). Political ideology and differences in seeking COVID-19 information on the internet: examining the comprehensive model of information seeking. *Online Information Review*.

Johnston, C. D. (2018). Authoritarianism, affective polarization, and economic ideology. *Political Psychology*, *39*, 219–238.

Jones, C. (2023). 'We the people, not the sheeple': Qanon and the transnational mobilisation of millennialist far-right conspiracy theories. *First Monday*.

Kamalakkannan, R., Rajkumar, R., Raj, M. M., & Devi, S. S. (2014). Imagined speech classification using eeg. *Advances in Biomedical Science and Engineering*, *1*(2), 20–32.

Kaufman, M., Kaufman, S., & Diep, H. T. (2022). Statistical mechanics of political polarization. *Entropy*, *24*(9), 1262.

Kaye, B. K., & Johnson, T. J. (2016). Across the great divide: How partisanship and perceptions of media bias influence changes in time spent with media. *Journal of Broadcasting & Electronic Media*, *60*(4), 604–623.

Keele, S., et al. (2007). Guidelines for performing systematic literature reviews in software engineering.

Kong, Q., Booth, E., Bailo, F., Johns, A., & Rizoiu, M.-A. (2022). Slipping to the extreme: A mixed method to explain how extreme opinions infiltrate online discussions. In *Proceedings of the international AAAI conference on web and social media*: *vol. 16*, (pp. 524–535).

Kovacs, E.-R., Cotfas, L.-A., & Delcea, C. (2022). From unhealthy online conversation to political violence: The case of the january 6th events at the capitol. In *International conference on computational collective intelligence* (pp. 3–15). Springer.

Kumar, S., Ravi, K., Mulay, S., Ram, K., & Sivaprakasam, M. (2018). Deep residual network based automatic image grading for diabetic macular edema. In *Research poster papers of the 2018 40th annual international conference of the IEEE engineering in medicine and biology society*. IEEE, http://dx.doi.org/10.13140/RG.2.2.24611.02082/1, URL https://www.researchgate.net/publication/374471910_Deep_Residual_Network_based_Automatic_Image_Grading_for_Diabetic_Macular_Edema.

Lahoti, P., Garimella, K., & Gionis, A. (2018). Joint non-negative matrix factorization for learning ideological leaning on Twitter. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 351–359).

Le, H. T., Boynton, G., Mejova, Y., Shafiq, Z., & Srinivasan, P. (2017). Revisiting the american voter on Twitter. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 4507–4519).

Le, H., Shafiq, Z., & Srinivasan, P. (2017). Scalable news slant measurement using Twitter. In *Proceedings of the international AAAI conference on web and social media*: *vol. 11*, (no. 1), (pp. 584–587).

Lee, T.-T. (2005). The liberal media myth revisited: An examination of factors influencing perceptions of media bias. *Journal of Broadcasting & Electronic Media*, *49*, 43.

Lee, Y., & Pirim, H. (2023). Comparison of parler and Twitter data using NLP: US capitol incident. In *IIE annual conference. proceedings* (pp. 1–6). Institute of Industrial and Systems Engineers (IISE).

Liang, C. S. (2022). Far-right contagion: the global challenge of transnational extremist networks. In *Handbook of security science* (pp. 1001–1037). Springer.

Linvill, D. L., Boatwright, B. C., Grant, W. J., & Warren, P. L. (2019). "The Russians are hacking my brain!" investigating Russia's internet research agency Twitter tactics during the 2016 United States presidential campaign. *Computers in Human Behavior*, *99*, 292–300.

Liu, R., Jia, C., Wei, J., Xu, G., & Vosoughi, S. (2022). Quantifying and alleviating political bias in language models. *Artificial Intelligence*, *304*, Article 103654.

Luttig, M. D. (2017). Authoritarianism and affective polarization: A new view on the origins of partisan extremism. *Public Opinion Quarterly*, *81*(4), 866–895.

Mahata, D., Zhang, H., Uppal, K., Kumar, Y., Shah, R., Shahid, S., et al. (2019). MIDAS at SemEval-2019 task 6: Identifying offensive posts and targeted offense from Twitter. In *13th international workshop on semantic evaluation* (pp. 683–690).

Malouf, R., & Mullen, T. (2008). Taking sides: User classification for informal online political discourse. *Internet Research*, *18*(2), 177–190.

Manickam, I., Lan, A. S., Dasarathy, G., & Baraniuk, R. G. (2019). IdeoTrace: a framework for ideology tracing with a case study on the 2016 US presidential election. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 274–281).

Martel, C., Allen, J., Pennycook, G., & Rand, D. G. (2022). Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, Article 17456916231190388.

Mason, L., & Wronski, J. (2018). One tribe to bind them all: How our social group attachments strengthen partisanship. *Political Psychology*, *39*, 257–277.

Matias, S. M. M., Costales, J. A., & Christian, M. (2022). A framework for cybercrime prediction on Twitter tweets using text-based machine learning algorithm. In *2022 5th international conference on pattern recognition and artificial intelligence* (pp. 235–240). IEEE.

Maynard, D., & Funk, A. (2012). Automatic detection of political opinions in tweets. In *The semantic web: ESWC 2011 workshops: ESWC 2011 workshops, Heraklion, Greece, May 29-30, 2011, revised selected papers 8* (pp. 88–99). Springer.

Medina Serrano, J. C., Papakyriakopoulos, O., & Hegelich, S. (2020). Dancing to the partisan beat: A first analysis of political communication on TikTok. In *Proceedings of the 12th ACM conference on web science* (pp. 257–266).

Melton, J., Bagavathi, A., & Krishnan, S. (2020). DeL-haTE: a deep learning tunable ensemble for hate speech detection. In *2020 19th IEEE international conference on machine learning and applications* (pp. 1015–1022). IEEE.

Morgan, J. S., Lampe, C., & Shafiq, M. Z. (2013). Is news sharing on Twitter ideologically biased? In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 887–896).

Morris, D. S., Morris, J. S., & Francia, P. L. (2020). A fake news inoculation? Fact checkers, partisan identification, and the power of misinformation. *Politics, Groups, and Identities*, *8*(5), 986–1005.

Neo, R. L. (2021). Linking perceived political network homogeneity with political social media use via perceived social media news credibility. *Journal of Information Technology & Politics*, *18*(3), 355–369.

Ness, E., Fatima, A., & Oghaz, M. M. D. (2023). Data driven model to investigate political bias in mainstream media. *IEEE Access*.

Nguyen, H., & Gokhale, S. (2022). An efficient approach to identifying anti-government sentiment on Twitter during Michigan protests. *PeerJ Computer Science*, *8*, Article e1127.

Noel, H. (2016). Ideological factions in the republican and democratic parties. *The Annals of the American Academy of Political and Social Science*, *667*(1), 166–188.

O'Hara, K., & Stevens, D. (2015). Echo chambers and online radicalism: Assessing the internet's complicity in violent extremism. *Policy & Internet*, *7*(4), 401–422.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13.

Olteanu, A., Cernian, A., & Gâgă, S.-A. (2022). Leveraging machine learning and semi-structured information to identify political views from social media posts. *Applied Sciences*, *12*(24), 12962.

Owoeye, K. O., & Weir, G. R. S. (2018). Classification of radical web text using a composite-based method. In *2018 international conference on computational science and computational intelligence* (pp. 53–58). http://dx.doi.org/10.1109/CSCI46756.2018.00018.

Pennacchiotti, M., & Popescu, A.-M. (2011). Democrats, republicans and starbucks afficionados: user classification in Twitter. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 430–438).

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, *64*, 1–18.

Prakasam, N., & Huxtable-Thomas, L. (2021). Reddit: Affordances as an enabler for shifting loyalties. *Information Systems Frontiers*, *23*, 723–751.

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond binary labels: Political ideology prediction of Twitter users. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 729–740).

Qi, X., Christensen, K., Duval, R., Fuller, E., Spahiu, A., Wu, Q., et al. (2010). A hierarchical algorithm for clustering extremist web pages. In *2010 international conference on advances in social networks analysis and mining* (pp. 458–463). IEEE.

Rajendran, A., Sahithi, V. S., Gupta, C., Yadav, M., Ahirrao, S., Kotecha, K., et al. (2022). Detecting extremism on Twitter during U.S. capitol riot using deep learning techniques. *IEEE Access*, *10*, 133052–133077.

Ramaciotti Morales, P. (2022). Multidimensional online American politics: Mining emergent social cleavages in social graphs. In *International conference on complex networks and their applications* (pp. 176–189). Springer.

Ravi, K., Selvaraj, S., Mulay, S., Ram, K., & Sivaprakasam, M. (2018). Breast cancer histology classification using deep residual networks. In *Research poster papers of the 2018 40th annual international conference of the IEEE engineering in medicine and biology society*. IEEE, http://dx.doi.org/10.13140/RG.2.2.22094.43840, URL https://www.researchgate.net/publication/374471906_Breast_cancer_histology_classification_using_Deep_Residual_Networks.

Ravi, K., & Vela, A. E. (2024a). Comprehensive dataset of user-submitted articles with ideological and extreme bias from reddit. *Data in Brief*, Article 110849. http://dx.doi.org/10.1016/j.dib.2024.110849, URL https://www.sciencedirect.com/science/article/pii/S2352340924008138.

Ravi, K., & Vela, A. E. (2024b). RICo: Reddit ideological communities. *Online Social Networks and Media*, *42*, Article 100279. http://dx.doi.org/10.1016/j.osnem.2024.100279, URL https://www.sciencedirect.com/science/article/abs/pii/S2468696424000041?via%3Dihub.

Ravi, K., Vela, A. E., & Ewetz, R. (2022). Classifying the ideological orientation of user-submitted texts in social media. In *2022 21st IEEE international conference on machine learning and applications* (pp. 413–418). IEEE, http://dx.doi.org/10.1109/ICMLA55696.2022.00066, URL https://ieeexplore.ieee.org/document/10069289.

Ravi, K., Vela, A. E., Jenaway, E., & Windisch, S. (2023). Exploring multi-level threats in telegram data with AI-human annotation: A preliminary study. In *2023 22nd IEEE international conference on machine learning and applications*. IEEE, http://dx.doi.org/10.1109/ICMLA58977.2023.00229, URL https://ieeexplore.ieee.org/abstract/document/10459792.

Ravi, K., & Yuan, J.-S. (2024). ThreatGram 101: Extreme telegram replies data with threat levels. In J. Lossio-Ventura, & et al. (Eds.), *Information Management and Big Data. SIMBig 2024. Communications in Computer and Information Science*. Springer, Cham, Just accepted.

Resnick, P., Alfayez, A., Im, J., & Gilbert, E. (2023). Searching for or reviewing evidence improves crowdworkers' misinformation judgments and reduces partisan bias. *Collective Intelligence*, *2*(2), Article 26339137231173407.

Ribeiro, F., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., et al. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proceedings of the international AAAI conference on web and social media*: *vol. 12*, (no. 1).

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., & Meira, W. (2020). Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 131–141).

Rudinac, S., Gornishka, I., & Worring, M. (2017). Multimodal classification of violent online political extremism content with graph convolutional networks. In *Thematic workshops '17, Proceedings of the on thematic workshops of ACM multimedia 2017* (pp. 245–252). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3126686.3126776.

Saez-Trumper, D., Castillo, C., & Lalmas, M. (2013). Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on information & knowledge management* (pp. 1679–1684).

Sarkar, R. (2022). An analytical approach for reducing k-line failure analysis and load shed computation. *IET Generation, Transmission & Distribution*, *16*(13), 2623–2641.

Shaw, A., & Benkler, Y. (2012). A tale of two blogospheres: Discursive practices on the left and right. *American Behavioral Scientist*, *56*(4), 459–487.

Shi, Y., Mast, K., Weber, I., Kellum, A., & Macy, M. (2017). Cultural fault lines and political polarization. In *Proceedings of the 2017 ACM on web science conference* (pp. 213–217).

Simas, E. N. (2018). Ideology through the partisan lens: Applying anchoring vignettes to US survey research. *International Journal of Public Opinion Research*, *30*(3), 343–364.

Simons, B., & Skillicorn, D. B. (2020). A bootstrapped model to detect abuse and intent in white supremacist corpora. In *2020 IEEE international conference on intelligence and security informatics* (pp. 1–6). IEEE.

Sipka, A., Hannak, A., & Urman, A. (2022). Comparing the language of QAnon-related content on Parler, Gab, and Twitter. In *Proceedings of the 14th ACM web science conference 2022* (pp. 411–421).

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437.

Spies, J. R. (2013). *The open science framework: improving science by making it open and accessible*. University of Virginia.

Sterling, J., Jost, J. T., & Hardin, C. D. (2019). Liberal and conservative representations of the good society: A (social) structural topic modeling approach. *Sage Open*, *9*(2), Article 2158244019846211.

Stier, S. (2016). Partisan framing of political debates on Twitter. In *Proceedings of the 8th ACM conference on web science* (pp. 365–366).

Swann, T., & Husted, E. (2017). Undermining anarchy: Facebook's influence on anarchist principles of organization in Occupy Wall Street. *The Information Society*, *33*(4), 192–204.

Tamer, M., Khamis, M. A., Yahia, A., Khaled, S., Ashraf, A., & Gomaa, W. (2023). Arab reactions towards Russo-Ukrainian war. *EPJ Data Science*, *12*(1), 36.

Tien, J. H., Eisenberg, M. C., Cherng, S. T., & Porter, M. A. (2020). Online reactions to the 2017 'unite the right'rally in charlottesville: measuring polarization in Twitter networks using media followership. *Applied Network Science*, *5*(1), 1–27.

Tran, M. (2020). How biased are American media outlets? A framework for presentation bias regression. In *2020 IEEE international conference on big data* (pp. 4359–4364). IEEE.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, *14*(3), 207–222.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of Internal Medicine*, *169*(7), 467–473.

Walther, S., & McCoy, A. (2021). US extremism on Telegram. *Perspectives on Terrorism*, *15*(2), 100–124.

Wang, T., Wang, K. C., Erlandsson, F., Wu, S. F., & Faris, R. (2013). The influence of feedback with different opinions on continued user participation in online newsgroups. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 388–395).

Wang, Y., Zannettou, S., Blackburn, J., Bradlyn, B., De Cristofaro, E., & Stringhini, G. (2021). A multi-platform analysis of political news discussion and sharing on web communities. In *2021 IEEE international conference on big data* (pp. 1481–1492). IEEE.

Waseem, Z. (2016). Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138–142).

Withers, K. L., Parrish, J. L., Terrell, S., & Ellis, T. J. (2017). The relationship between the "dark triad" personality traits and deviant behavior on social networking sites.

Wojcieszak, M., Winter, S., & Yu, X. (2020). Social norms and selectivity: Effects of norms of open-mindedness on content selection and affective polarization. *Mass Communication and Society*, *23*(4), 455–483.

Wong, M. A., Frank, R., & Allsup, R. (2015). The supremacy of online white supremacists–an analysis of online discussions by white supremacists. *Information & Communications Technology Law*, *24*(1), 41–73.

Wong, F. M. F., Tan, C. W., Sen, S., & Chiang, M. (2016). Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Transactions on Knowledge and Data Engineering*, *28*(8), 2158–2172.

Xiao, Z., Zhu, J., Wang, Y., Zhou, P., Lam, W. H., Porter, M. A., et al. (2023). Detecting political biases of named entities and hashtags on Twitter. *EPJ Data Science*, *12*(1), 20.

Yang, M., & Chen, H. (2012). Partially supervised learning for radical opinion identification in hate group web forums. In *2012 IEEE international conference on intelligence and security informatics* (pp. 96–101). IEEE.

Youngblood, M. (2020). Extremist ideology as a complex contagion: the spread of far-right radicalization in the United States between 2005 and 2017. *Humanities and Social Sciences Communications*, *7*(1), 1–10.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666.

Zaytoon, M., Bashar, M., Khamis, M. A., & Gomaa, W. (2024). Amina: an Arabic multi-purpose integral news articles dataset. *Neural Computing and Applications*, 1–21.

Zerrer, P., & Engelmann, I. (2022). Users' political motivations in comment sections on news sites. *International Journal of Communication*, *16*, 23.

Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2018). A survey on deep learning for big data. *Information Fusion*, *42*, 146–157.

Zhu, X., Caliskan, C., Christenson, D. P., Spiliopoulos, K., Walker, D., & Kolaczyk, E. D. (2023). Disentangling positive and negative partisanship in social media interactions using a coevolving latent space network with attractors model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *186*(3), 463–480.