IMT School for Advanced Studies, Lucca
Lucca, Italy

Cybersecurity and Cyber Intelligence Measures for Monitoring, Preventing, and Mitigating Radicalization Pathways

PhD Program in Cybersicurezza

Track in Human, Economic, and Legal Aspects in Cybersecurity

XXXVIII Cycle

By

Omran Berjawi

2025

The dissertation of Omran Berjawi is approved.

PhD Program Coordinator: Rocco De Nicola, IMT School for Advanced Studies Lucca

Advisor: Prof. Giuseppe Fenza, University of Salerno

The dissertation of Omran Berjawi has been reviewed by:

Prof Sherali Zeadally, University of Kentucky

Hamid Mcheick, Univeristy of Quebec at Chicoutimi

IMT School for Advanced Studies Lucca
2025

Contents

List of Figures

List of Tables

Acknowledgements

I would also like to express my deepest gratitude to my advisor, Prof. Giuseppe Fenza, for his invaluable guidance and support throughout my Ph.D. journey. His insightful advice and patience have been indispensable to my growth, both as a researcher and as an individual. This endeavor would not have been possible without his belief in my potential and the opportunities he provided to help me achieve my goals.

Vita

Publications

1. O. Berjawi, G. Fenza, and V. Loia, "A comprehensive survey of detection and prevention approaches for online radicalization: Identifying gaps and future directions," in *IEEE Access*, vol. 11, pp. 120463-91, Oct. 2023.
2. O. Berjawi, D. Cavaliere, G. Fenza, and V. Loia, "Understanding radicalization pathways: a framework for assessing diversity in YouTube recommendation systems," in *Social Network Analysis and Mining*, vol. 14, no. 1, pp. 233, 2024.
3. O. Berjawi, G. Fenza, R. Khatoun, and V. Loia, "Mitigating radicalization in recommender systems by rewiring graph with deep reinforcement learning," in *Online Social Networks and Media*, vol. 48, pp. 100325, 2025.

Conferences / Presentations

1. O. Berjawi, D. Cavaliere, and G. Fenza, "A Multi-aspect Analysis of Echo Chambers on Video-Sharing Social Media," in *International Conference on Advances in Social Networks Analysis and Mining*, pp. 197-213, Cham: Springer Nature Switzerland, Sep. 2024.
2. O. Berjawi, R. Khatoun, W. Fahs, and G. Fenza, "Leveraging Sentiment and Emotion Analysis to Enhance Cyberbullying Detection," in *2024 11th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 81-86, IEEE, Dec. 2024.
3. O. Berjawi, D. Cavaliere, G. Fenza, and R. Khatoun, "Dynamic analysis of influencer impact on opinion formation in social networks," in *International Conference on Web Information Systems Engineering*, pp. 394-408, Singapore: Springer Nature Singapore, Dec. 2024.
4. O. Berjawi, R. Khatoun, and G. Fenza, "Digital Persuasion: Understanding the Impact of Online Influencers on Public Opinion," in *International Conference on Persuasive Technology*, pp. 117-127, Cham: Springer Nature Switzerland, May 2025.
5. O. Berjawi, R. Khatoun, and G. Fenza, " Analyzing the Persuasive Strategies of Influencers and News Media on Social Media". To appear in the International Conference on Computer Systems and Applications (AICCSA 2026).

Abstract

Social media platforms have reshaped the way users communicate, providing substantial benefits, but also introduced significant concerns about accelerating ideological enclosure and online radicalization through the combined effects of user behavior, algorithmic personalization, and persuasive actors. This thesis investigates these interlocking mechanisms by developing integrated detection, diagnosis, and mitigation methods that combine deep learning (DL), machine learning (ML), natural language processing (NLP), and network analysis to track radicalization pathways within online environments. Empirical evaluations on large, real-world datasets from video-sharing and microblogging platforms demonstrate three principal findings. First, when behavioral and affective signals are integrated with network representations, they improve the detection and forecasting of echo-chamber persistence, and they correlate strongly with emotional polarization and toxic interactions. Second, a recommender-diversity diagnostic identifies latent radicalization risk within recommendation graphs, and an adaptive graph-rewiring policy (DRLGR) reduces exposure to radicalizing pathways while preserving engagement metrics. Third, influential actors exert outsized effects on opinion dynamics: rhetorical alignment with audiences magnifies influence beyond direct followers. The thesis contributes (i) a systematic survey and gap analysis, (ii) a behavioral echo-chamber detection and forecasting framework, (iii) a diagnostic and learned mitigation method for recommender systems, and (iv) an integrated model of influencer impact that couples network position and rhetorical strategy.

Chapter 1

Introduction

The digital revolution has reshaped the lifestyles of billions of people worldwide, becoming a fundamental part of daily life that influences both personal and professional activities. An aspect of this transformation is the emergence of social networks (eg, Twitter, Facebook, and Instagram), whose popularity has grown remarkably in recent years, with a global user base growing from 3.19 billion in 2018 to a projected 5.20 billion by 2025 [95]. While social media platforms offer substantial benefits, such as overcoming traditional information barriers and promoting interaction among users from diverse backgrounds, they also present significant risks that cannot be overlooked. These platforms rely heavily on algorithmic systems that personalize the content to enhance users' engagement based on their preferences and prior interactions. However, this often results in the formation of echo chambers [60, 122]. This dynamic inadvertently fuels the phenomenon of online radicalization, as individuals gradually adopt extreme political, social, or religious views through prolonged interaction with content that reinforces their views and like-minded communities [119].

This mechanism leads these platforms to take actions to control this emerging phenomenon. Between 2020 and 2023, Meta, X (formerly Twitter), and YouTube intensified their efforts to remove content. In Q1 2020, Facebook removed around 4.7 million pieces of hate content, with 96.7% of it

flagged by AI [80]. During the first half of 2024, X suspended over 57,000 accounts tied to violent hate and removed 153 pieces of extremist content [111]. During the same period, X processed more than 67 million reports of hateful content, resulting in over 2,000 account suspensions [127]. In a separate 2022 audit, researchers identified 435 militia-affiliated videos still present on the YouTube platform, highlighting persistent hosting of extremist content despite takedown efforts [20]. In parallel, researchers have made significant contributions to understanding and addressing this phenomenon. For example, Sharma et al. [112] examined how recommendation algorithms on Twitter contribute to polarization, while other studies [2, 5] adopted approaches to study the impact of RS on user behavior. Daly et al. [31] found that friend recommendation algorithms intensified network clustering, reinforcing ideological homogeneity. Furthermore, ML and DL methods were used to detect extremist and hate speech content as a content-based intervention strategy [8, 4, 85].
However, despite these concerted efforts, the effectiveness of such interventions remains limited. The ongoing persistence and adaptability of online radicalization underscore the urgent need to prioritize updated strategies for monitoring and preventing this phenomenon. This thesis situates itself within this critical landscape by framing online radicalization as a cognitive security and information integrity threat, and advances methodologies at the intersection of cybersecurity and social media analysis to understand, track, and counter radicalization dynamics within digital environments.

1.1 Motivation

The online radicalization represents a serious threat to both digital security and societal stability; once considered a fringe concern, it has now emerged as a focal point on the international security agenda [36]. This phenomenon is exemplified by recent global events, which reveal how social media's role in spreading extremist content extends into influencing public discourse and undermining democratic institutions. During the COVID-19 crisis, platforms like YouTube and Facebook were ex-

ploited to propagate anti-vaccine rhetoric and conspiracy theories, ranging from virus denial to baseless claims about 5G technology and vaccine surveillance [131, 29, 90]. Such disinformation campaigns significantly eroded public trust in health authorities, especially given the growing dependence on online news sources [78]. Similarly, during the Brexit campaign and the 2020 U.S. presidential election, disinformation was deliberately amplified through algorithmic features that prioritized emotionally charged and misleading content, often promoted by opaque or unregulated actors [9, 52]. These incidents highlight the misuse of social media to manipulate public opinion on important global issues and underscore the urgency of addressing online extremism as both a sociopolitical and computational challenge.

In this context, the radicalization process is not random or isolated; instead, a synergistic relationship exists between algorithmic infrastructures and human behavior that shapes it. On one hand, recommendation algorithms designed to personalize content based on user preferences and behaviour can create self-reinforcing feedback loops that expose users to ideologically narrow content, fostering cognitive isolation [91, 25]. On the other hand, the communities that form within these echo chambers, along with influential actors such as leaders and online influencers, also play a crucial role in guiding users' ideological trajectories [11, 128]. These actors often employ emotionally resonant narratives and persuasive rhetoric to legitimize extremist viewpoints, and their influence is amplified by the very algorithmic systems that reward engagement. High-profile public figures, for instance, have demonstrated the significant impact of digital platforms on shaping political discourse. In the 2020 U.S. presidential election, celebrities like Dwayne Johnson and Taylor Swift actively encouraged voter participation and shared political endorsements, while in 2024, some celebrities endorsed Kamala Harris, like Taylor Swift and Jennifer Lopez, and some supported Donald Trump, such as Elon Musk [108, 28].

Thus, social media platforms have evolved from a medium of passive content consumption into an active ecosystem that reinforces ideological perspectives. Platform takedown and content-moderation efforts,

although increasingly prevalent, have had only partial success: problematic content persists and often re-emerges in novel forms. At the same time, purely content-centric approaches struggle to anticipate the structural and behavioral dynamics that lead to escalation. These observations highlight a significant gap. Existing research and intervention strategies are often siloed: network-structural analyses describe clustering; content classifiers detect abusive or extremist speech; and algorithmic audits reveal biases in recommender systems. However, few studies systematically combine behavioral signals, recommendation dynamics, and the agency of influential actors to produce operational tools for early detection and intervention. This thesis treats online radicalization as an inherently socio-technical problem—one that arises from the interplay among user behavior, algorithmic curation, and targeted rhetorical strategies and argues that effective monitoring and mitigation require integrated, proactive methods capable of (i) detecting emergent echo chambers, (ii) diagnosing algorithmic pathways that amplify ideological enclosure, and (iii) modelling how influential actors drive opinion shifts. The work that follows develops and evaluates such integrated methods using empirical social-media data.

1.2 Research Scope& Objectives

This thesis examines the structural, algorithmic, and human factors that drive online radicalization, with the dual objectives of enhancing detection and facilitating effective mitigation. The empirical scope focuses on data from large-scale online platforms, employing a combination of machine learning, network analysis, and natural language processing approaches. The thesis is organized around three research questions:

- **RQ1. How can echo chambers be detected and monitored in dynamic social-media environments?**
  The first research question focuses on enhancing the detection of echo chambers, which are recognized as one of the drivers of online radicalization. This thesis explores advanced methodologies for

identifying and monitoring these communities by moving beyond static network structures. Specifically, it aims to integrate dynamic behavioral indicators such as the linguistic patterns into detection models, thereby improving their accuracy and adaptability.

- **RQ2. Does the recommender system contribute to online radicalization?**
  This research question explores how RS, as those employed by YouTube, can foster ideological isolation and reinforce echo chambers. It further examines how redesigning these systems could serve as an effective countermeasure to mitigate such effects, preserving user autonomy and platform engagement.

- **RQ3. What role do influential users play in driving ideological change within online communities?**
  While much research has focused on algorithmic amplification and content-based drivers of online radicalization, the third research question addresses the human dimension of radicalization by examining the influence of prominent actors, including political figures, content creators, and social media influencers. This thesis models opinion dynamics and rhetorical strategies to investigate how emotionally charged and strategically framed messages shape public attitudes and ideological alignment. Particular attention is given to factors such as the credibility, network centrality, and persuasive tactics of these actors in amplifying and legitimizing radical content.

Each research question is addressed through a targeted empirical and methodological contribution (Chapters 3–5), evaluated on real-world datasets, and reported results.

1.3 Contributions

This thesis advances theory, methods, and practice in the study of online radicalization. The contributions are listed here in concise form and mapped to the research questions they address.

1.3.1 Contribution 1 — Survey and gap analysis.

This thesis provides a comprehensive review of approaches used to address online radicalization, drawing on approximately 70 peer-reviewed works. The survey examines a wide range of methodologies, including ML, DL, and graph-based techniques. It categorizes these approaches, highlights commonly used datasets, and explores the evolving relationship between social media platforms and extremist content. Furthermore, it identifies current limitations and offers actionable directions for future research.

- Classifies existing approaches for detecting and preventing online radicalization.
- Analyzes the datasets used in radicalization studies.
- Identifies major limitations, open challenges, and future research directions.

1.3.2 Contribution 1 — Detection and monitoring of echo chambers and toxic behaviors.

This thesis is motivated by a comprehensive survey of approximately 70 peer-reviewed studies on online radicalization, echo chambers, and harmful online behavior, which identifies key gaps in existing approaches, particularly their reliance on static network structures and limited consideration of temporal, behavioral, and affective dynamics. These findings directly inform the design of the first contribution, which provides a dynamic and behavior-aware approach to detecting and monitoring echo chambers in online social platforms.

The proposed framework incorporates behavioral indicators to capture users' linguistic patterns and interactions over time. The empirical study, conducted on YouTube data, demonstrates the effectiveness of these indicators at both the community and individual levels. First, the framework identifies echo chambers and tracks users' behavioral shifts as proxies for the temporal evolution of these communities. Second, at the individual level, the predictive model leverages behavioral indicators to forecast whether a user is likely to remain in or exit an echo chamber. Additionally, the analysis of toxic language within these communities reveals a strong correlation between emotional polarization and cyberbullying behavior among users in radicalized groups. Overall, this contribution provides a proactive tool for anticipating and monitoring the dynamic nature of online radicalized communities. Specifically, it:

- Introduces behavioral indicators as dynamic proxies for detecting and monitoring the evolution of echo chambers over time.
- Develops a predictive model to forecast user retention or departure from echo chambers based on behavioral data.
- Demonstrates significant correlations between toxic language, emotional polarization, and cyberbullying within radicalized online communities.

1.3.3 Contribution 2 — Diagnosis and mitigation of algorithmic radicalization in recommender systems.

Following the identification of echo chambers, this contribution examines the role of RS in facilitating online radicalization and proposes a mitigation strategy to reduce their impact. In this context, an indicator was designed to measure the degree to which users may be exposed to radicalization through YouTube RS. The empirical analysis demonstrates the effectiveness of this indicator, showing that radicalized recommendations are not randomly generated but are significantly influenced by specific video attributes and individual user behavior patterns. Consequently, a method was developed to rewire the recommendations for

the radicalization degree indicator dynamically. The findings show that DRLGR effectively increases diversity within recommendation flows, thereby reducing the overall level of radicalization. As such, this contribution provides a dual-purpose framework that serves both to measure radicalization risk and to implement proactive interventions when elevated risks are detected. This contribution is threefold:

- Proposes a diversity-based indicator to quantify the degree of radicalization within recommendation systems.
- Demonstrates how prior user activity and content attributes influence exposure to radical content.
- Introduces a method to mitigate radicalization by increasing ideological diversity in recommendations.

1.3.4 Contribution 3 — Modelling influencer impact and rhetorical strategies in opinion dynamics.

Beyond algorithmic structures and content-based indicators, this contribution investigates the role of influential digital actors in shaping public opinion and driving ideological shifts within online communities. To this end, the Friedkin–Johnsen (FJ) model was employed, capitalizing on its robustness in modeling opinion dynamics. First, the model demonstrates its effectiveness in identifying genuinely influential users who exert meaningful influence over their audiences beyond traditional metrics such as centrality measures. Second, the results indicate that these influential users have the capacity to significantly alter public opinion, with even minor shifts in their viewpoints potentially triggering wide-ranging effects across the community. Furthermore, a rhetorical analysis was conducted for both influencers and media actors, revealing that influencers often employ a hybrid strategy that combines credibility (ethos) with logical appeals (logos) to engage their audiences effectively. Consequently, this contribution demonstrates insights into how digital influencers can act as either accelerants or moderators of ideological polarization. The core contributions of this analysis are outlined below:

- Develops influence-based ranking systems that outperform traditional centrality metrics.
- Simulates opinion formation and change based on influence and exposure.
- Analyzes rhetorical strategies (ethos, pathos, logos) used by influencers and news media to shape opinion.

1.3.5 List of publications

The following peer-reviewed publications underpin and inform the research contributions of this thesis. Each publication corresponds to a key contribution discussed in this work:

Publication I: This publication presents a comprehensive survey that provides a thorough investigation into current computational methods for detecting online radicalization across various platforms. This analysis systematically explores the ML, graph-based, and behavioral detection techniques that have been applied while also identifying major limitations in the current literature. The study proposes a unified framework to guide future detection strategies, emphasizing the importance of context-aware models.

Publication II: In this publication, a framework was developed for detecting and analyzing the echo chambers by combining graph neural networks (GNNs) and NLP. The framework defines behavioral indicators based on user interactions to identify echo chambers and track their development over time. The validity of the proposed framework in monitoring is demonstrated; experiments show that these indicators correlate with established metrics at over 90% and can predict whether users will remain in or leave their communities. This confirms the effectiveness of behavior-based indicators in monitoring and forecasting echo chamber dynamics on video-sharing platforms.

Publication III: This work proposes a model that integrates TF-IDF with sentiment and emotion scores to detect the types of cyberbullying published online. The analysis further reveals distinct emotional patterns associated with different categories of cyberbullying, with negative emotions such as anger, disgust, and fear being predominantly linked to cyberbullying content. In contrast, non-cyberbullying content displayed a more balanced emotional profile, exhibiting higher values for neutral and positive emotions. These findings underscore the significant role of emotional and sentiment analysis in enhancing the detection of harmful behaviors in online environments.

Publication IV: A framework is introduced that combines XAI techniques with predictive modeling to investigate the role of YouTube's recommendation system (RS) in shaping radicalization pathways. The aim is to assess diversity levels as a proxy for measuring the increasingly radicalized content that users are exposed to through YouTube's up-next recommendations. To achieve this, an indicator is introduced to measure the diversity level within recommendation content, and the impact of video attributes and user history on the variation of diversity levels is studied. Experimental results show that the diversity score correlates with radicalization measures, indicating that specific video features and prior user history significantly influence the selection of radicalized up-next video recommendations. This contribution provides a valuable indicator for measuring the presence of radicalization within recommendation systems.

Publication V: This paper addresses the mitigation of radicalization within RS by introducing a framework based on graph theory to measure and mitigate ideological reinforcement. The proposed method defines a radicalization score *Rad(G)*, which measures the extent to which a user is stuck on a radical path within the recommendation graph. A proposed approach called DRLGR is then applied to iteratively adjust the graph structure by selecting and modifying recommendation edges that contribute to high radicalization exposure. Through continuous

learning, the model identifies rewiring strategies that lower *Rad(G)* while preserving recommendation utility. Experiments conducted on two different datasets reveal the effectiveness of DRLGR compared to baseline and heuristic approaches, as it achieves more sustained reductions in radicalization over time, particularly in complex network structures where traditional interventions tend to plateau.

Publication VI: This contribution presents a framework that integrates the FJ opinion dynamics model with sentiment analysis to study how influencers shape opinion change over time in social networks. Interaction networks are constructed, communities are detected, and key influencers are identified using structural centrality metrics. By injecting real-time sentiment values into the model, the simulation examines how opinions spread under the influence of highly influential users. The results reveal that influencers play a significant role not only in shifting opinions within their direct communities but also in influencing peripheral users beyond their immediate network. This study presents a dynamic and fine-grained approach for measuring the evolving influence of influencers on public sentiment in online platforms.

Publication VII: The FJ model is utilized to assess the impact of key actors on opinion dynamics during political discourse on social media. By simulating targeted manipulations of initial opinions, the difference in community opinion shifts is assessed when central influencers are perturbed versus random users. Using real-world Twitter data, it is demonstrated that top influencers identified by their influence scores can significantly influence the collective opinion of the network. Notably, the impact of influencers extends beyond direct followers, reaching secondary connections, underscoring the critical role they play in shaping and amplifying online political narratives.

Publication VIII: This study examines how political influencers and news media employ rhetorical persuasion to influence user engagement during political events, drawing on Aristotle's framework of ethos,

Table 1: Summary of Thesis Contribution Areas and Their Corresponding Publications

Contribution Area	**Publications**
Detection and monitoring of echo chambers and toxic behaviors	I, II, III
Diagnosis and mitigation of algorithmic radicalization in recommender systems	IV, V,
Modelling influencer impact and rhetorical strategies in opinion dynamics	VI, VII, VIII

pathos, and logos. The linguistic strategies used by each actor type and their correlation with audience reactions are analyzed. The findings reveal that influencers utilize a hybrid of credibility and logical appeals, exhibiting high linguistic homophily with their audiences. At the same time, news media rely predominantly on ethos-driven messaging with lower alignment. This contribution highlights the divergent communication strategies employed by influencers versus institutional actors and reveals how rhetorical adaptation enhances the persuasive impact of digital political discourse.

1.4 Thesis Organization

The structure of this thesis is as follows: Chapter 2 reviews the literature on approaches used to counter online radicalization. Chapter 3 presents two works that focus on detecting the presentness and the evolution of the echo chamber. Chapter 4 introduces the recommendation-diversity diagnostic and the DRLGR graph-rewiring mitigation method. Chapter 5 presents the work that examines the role of influential actors. Chapter 6 synthesizes the findings, discusses implications, and outlines future research directions.

Chapter 2

Literature Review

Monitoring and mitigating online radicalization is addressed in various research communities, spanning political science, sociology, social networking, and data mining. An extensive survey of research articles sourced from IEEE, Scopus, and Web of Science was conducted, covering the period from 2017 to 2023. The methodology employed a combination of keyword-based Boolean queries and a systematic filtering process. The review begins by examining studies that focus on detecting echo chambers in social networks. It then provides an overview of research examining the role of recommendation systems (RS) in online radicalization. Finally, it presents current efforts aimed at investigating influential actors in opinion dynamics.

2.1 Empirical studies of echo chambers

Understanding the role of the echo chamber in political polarization, misinformation spread, and the fragmentation of public discourse has become a significant concern. Social networks inherently encourage homophily and selective exposure, fostering environments where users may become insulated from opposing viewpoints and thereby reinforce their existing beliefs. The empirical study of echo chambers has thus evolved into a multidisciplinary effort, employing diverse method-

ologies including social network analysis, content-based approaches, information-theoretic measures, and graph-based modelling to detect, quantify, and explain their presence. These studies have analyzed various social networks and topical domains (e.g., elections, COVID-19, conspiracy theories), revealing that a combination of user behavior, algorithmic curation, and network structures shapes the formation of echo chambers. While some research focuses on specific case studies to trace how echo chambers form and evolve around controversial topics, others propose generalizable metrics or models grounded in network topology or user interaction patterns. Collectively, these empirical approaches shed light on both the structural characteristics and social mechanisms that sustain echo chambers in digital environments.

A significant contribution to echo chamber quantification is offered by [3], who propose a score called ECS that assesses ideological cohesion and separation among users in embedding space. The innovation lies in the use of a self-supervised graph autoencoder (EchoGAE) to learn embeddings reflective of ideological positions without requiring explicit labels or assumptions about network topology. Their findings affirm ECS as an effective score supporting the argument that embedding-based techniques can be leveraged for scalable echo chamber analysis. Complementing this metric-based quantification, Minici et al. [81] introduce a generative probabilistic framework that jointly models user network structure and information diffusion to uncover latent ideological communities. Its strength lies in inferring both the degree of ideological alignment and the level of insulation from counter-ideological content. This work underscores the interplay between user connectivity and information flow in the materialization of echo chambers.

Another topological metric, Random Walk Controversy (RWC), was introduced by Garimella et al. [47]. This measure quantifies the probability of a random walk originating from a community and terminating in a different community, thereby capturing the degree of separation between communities in a network. Similarly, Diaz et al [35] examined information diffusion across communities to assess the presence and strength of echo chambers, using inter-community information flow as a key indica-

tor. Kratzke et al. [69] adopts a structurally similar approach, employing graph-based detection on retweet interactions in the German-language Twitter sphere. Using modularity and HITS algorithms, the study identifies ideologically cohesive echo chambers, highlighting the role of orchestrated trolls in reinforcing ideological silos through purely interactional signals.

Other studies have taken a more application-specific route. Villa et al. [121] integrating topological and semantic signals from Twitter interactions to study echo chambers. By applying community detection techniques to both structural and content-aware graph representations, they reveal how pandemic discourse facilitated the formation of polarized clusters. Notably, their work affirms the role of semantic homogeneity in reinforcing echo chambers, a point corroborated by earlier research emphasizing selective exposure and homophily in information environments [46, 63]. In line with this focus on pandemic discourse, Di Marco et al. [34] examine infodemics on YouTube by analyzing 13,000 videos and over 2 million user engagements. The study finds clear evidence of echo chambers structured by both political bias and information reliability, with such clustering disappearing under randomization—affirming the socially constructed nature of these silos.

Efforts toward standardization in echo chamber detection are exemplified by [83], who propose a four-stage framework applicable across platforms. Their case study on Reddit examines debates from the early years of Donald Trump's presidency, highlighting the persistence and temporal stability of ideological communities. Interestingly, findings suggest that ideological polarization does not always equate to total insulation, a nuance often overlooked in more binary treatments of echo chambers. This observation resonates with [53], who adopts a cultural lens to examine conspiracy communities on YouTube. Through mixed-methods analysis of user comments, the study reveals that echo chamberness varies significantly based on subcultural values, suggesting that insularity is not uniformly distributed but contingent upon community norms and interpretive frames.

Echo chamber research has also expanded into content-centric method-

ologies. Calder et al. [19] focus on detecting echo chambers by analyzing the stance and emotional alignment between posts and comments. By designing semi-supervised features to capture ideological and affective resonance, they quantify the degree of echoing in discourse. This work highlights an important dimension often ignored in network-based models: the affective and rhetorical congruity that binds users within ideological enclaves. Sun et al. [114] provide further insight into comment-based interactions, studying YouTube vaccine video discussions through a contagion model informed by social interaction patterns. The study identifies a weak echo chamber effect reinforced by imitation and intergroup interactions, with reciprocity playing a lesser role—underscoring how local interaction rules shape the spread of health misinformation. Ghafouri et al. [50] introduce a Transformer-based framework to quantify the echo chamber effect through user diversity and separability. Their unsupervised model integrates linguistic and behavioral data, yielding a nuanced metric that can detect polarization and homogeneity across different ideological groups.

Further comparative insight is provided by [65], who evaluate detection metrics across Reddit and Twitter. Their cross-platform analysis reveals that the performance and interpretability of echo chamber metrics vary significantly depending on platform affordances and community structures. Echo chamber effects in newer digital environments are exemplified by [45], who study short video platforms, such as TikTok. By analyzing selective exposure and homophily in both user and algorithmic recommendations, they highlight how short-form video content accelerates ideological clustering. Interestingly, their findings suggest that cultural and platform-specific moderators can either exacerbate or mitigate the formation of echo chambers.

Finally, broader questions of polarization and radicalization, which are intrinsically linked to echo chamber dynamics, have been addressed in [84] and [49]. Muñoz introduces a novel algorithm to capture political polarization during election cycles in Spain, demonstrating how political discourse becomes increasingly fragmented over time. Ghafouri further expands this line of inquiry by applying large language models (LLMs)

to measure ideological diversity and track radicalization across platforms and issue domains. These studies collectively demonstrate that echo chambers not only reflect ideological segregation but also catalyze shifts toward extremity, reinforcing their centrality in the study of digital political dynamics. In summary, empirical research on echo chambers has evolved from simple network homophily analyses to more complex models that incorporate topology, content, semantics, and user behavior. While early studies focused on detecting ideological clustering, contemporary approaches increasingly emphasize quantification, cross-platform comparability, and the integration of advanced ML techniques. Together, these works illustrate that echo chambers are not merely artifacts of social connectivity, but emergent phenomena shaped by a complex interplay of algorithms, content, and human agency.

2.1.1 Role of Algorithmic Recommendation Systems in Amplifying Radical Pathways

The potential for algorithmic RS to shape ideological exposure and reinforce radical beliefs has been a major focus in recent literature. A key concern is the creation of radicalization pathways, wherein users are gradually steered toward more extreme content through algorithmic curation. This concern has been investigated across multiple platforms and via diverse methodological approaches.

Early large-scale audits such as [98] and [73] demonstrated that YouTube's RS frequently guided users toward increasingly extreme right-wing content, thereby reinforcing ideological silos. Haroon et al. [58] extended this work by auditing ideological bias within YouTube's algorithm and proposing interventions to mitigate radicalization. In a follow-up study [57], a more detailed analysis of ideological and problematic content exposure was conducted. In contrast, Srba et al. [110] examined the formation of misinformation filter bubbles resulting from algorithmic loops. Perra et al. [93] proved that a graph can be used to find some key drivers of the emergence of an echo chamber, like the triangles. Similar dynamics have been observed on other platforms. Shin

et al. [107] reverse-engineered TikTok's recommendation engine, showing that the platform's design can compress radicalization timelines by quickly exposing users to ideologically extreme content. Tommasel et al. [116] further emphasized the vulnerability of engagement-optimized systems to misinformation super-spreaders.

Beyond empirical audits, computational models offer a theoretical foundation for understanding these effects. Lanzetti et al. [71] employed agent-based simulations to demonstrate how RS can simultaneously drive both individual-level and societal polarization. Bellina et al. [10] focused specifically on collaborative filtering algorithms, demonstrating that even mainstream recommender techniques can amplify polarization under specific network structures and feedback dynamics. Daly et al. [30] used the similarity of interest between friends to reveal the impact of friends' recommendations.

In response to the risks posed by algorithmic radicalization, a range of mitigation strategies has been proposed. Fabbri et al. [39] explored the use of graph rewiring to adjust YouTube's watch-next recommendations, demonstrating that minor structural changes can reduce radical content exposure while maintaining relevance. Coupette et al. [27] confirmed that even minimal rewiring of recommendation graphs can significantly curb exposure to harmful content without compromising user satisfaction. Similarly, Gao et al. [43] introduced CIRS, a counterfactual RS that strategically exposes users to ideologically dissimilar but credible content to disrupt filter bubbles.

RL has emerged as another promising approach. Li et al. [74] proposed a controllable RS that explicitly balances personalization and ideological diversity using RL. Geng et al. [48] developed a bio-inspired algorithm to optimize the accuracy-diversity tradeoff in recommendation, showing that it increases exposure to diverse content. Wang et al. [125] introduced a graph-based nudging mechanism that strategically alters user trajectories within the recommendation graph to counter belief reinforcement and filter bubbles.

Network-based interventions also show potential. Haddadan et al. [55]

proposed Repbublik, a method that adds targeted links to reduce the effective radius of polarized communities, thereby encouraging cross-ideological exposure. De Arruda et al. [6] presented a dynamic model illustrating how recommendation-induced network rewiring can intensify polarization, and proposed structural interventions. Cinus et al. [26] studied people-recommenders (suggestions for who to follow rather than what to watch), showing that poorly calibrated systems can also exacerbate echo chambers.
Within the domain of news consumption, Vercoutere et al. [120] proposed a hybrid graph-based RS that balances selection diversity and relevance, specifically for politically sensitive content. Rahman et al. [97] developed a stochastic dynamic programming model that optimizes article recommendation to favor cross-cutting ideological exposure while preserving user engagement. Santos et al. [104] evaluated link recommendation algorithms and their influence on the polarization trajectories in social networks, finding that small changes in connectivity can substantially alter ideological exposure over time.

Taken together, these studies reveal a multifaceted picture of how RS can influence belief formation, ideological segregation, and radicalization. Empirical audits provide concrete evidence of harmful exposure patterns, while computational models and algorithmic experiments underscore the structural roots of these problems. At the same time, numerous intervention strategies, ranging from graph rewiring and belief-aware systems to RL and hybrid recommender design, show promise in mitigating these effects. As the influence of RS continues to grow, building transparency, diversity, and responsibility into algorithmic design should remain a central goal of both researchers and platform developers.

2.2 Influential Actors in Opinion Dynamics

The dynamics of opinion, influencer identification, and the societal impact of digital persuasion are deeply intertwined. Understanding how

individuals form and adapt their views, especially in fragmented and polarized online communities, requires a multidisciplinary approach that connects the structural features of social networks with the behaviors of influential agents and the mechanisms through which they influence discourse.

2.2.1 Understanding Digital Persuasion and Influence Dynamics

Models grounded in opinion dynamics, such as the FJ model, have been widely used to explore these mechanisms. The FJ model and its extensions account for agents with fixed opinions commonly referred to as stubborn agents who resist updates from others [115]; [136]. Recent works have expanded these frameworks to multi-dimensional settings (e.g., agents balancing opinions across different issues), showing that a small number of such stubborn agents can prevent convergence even under wide exposure. Further extending this analysis, researchers have introduced signed graphs and dynamic edge models [137], [16] to capture both positive and negative relationships (e.g., trust versus distrust or agreement versus disagreement) as well as evolving connections. These models reveal that edge rewiring, such as unfollowing dissenting voices or reinforcing in-group ties, accelerates polarization and the formation of echo chambers. At a macro level, opinion optimization approaches, such as those by Sun et al. [113], examine how intervention—typically by strategically placing persuasive agents — can shift global opinion distributions. These interventions often leverage submodular optimization, enabling targeted disruption of polarized structures.

Together, these studies illustrate that echo chambers are not fixed phenomena but rather emergent patterns arising from network topology, interpersonal influence, and resistance to opinion change. The interplay between local dynamics (individual stubbornness or tie strength) and global structure (community clustering) is crucial for understanding both how polarization emerges and how it can be mitigated.

2.2.2 Identification of Influencers and Political Leaders

Detecting influential actors in social networks is central to understanding how opinions are shaped and propagated, particularly in the context of political discourse and polarization. A wide range of methods have emerged to identify these actors, often combining topological metrics with contextual and behavioral analysis. Several works propose enhanced centrality measures that consider community structure, structural holes, and propagation dynamics. Zhang et al. [134] and Zhao et al. [135] demonstrate that influence is more accurately measured not only by a node's direct connections but also by its ability to bridge communities and occupy structural gaps. More recent approaches incorporate semi-local metrics [133], multi-scale propagation strength [54], and random-walk distance-awareness [23] to better model realistic spreading behavior.

Beyond structural factors, researchers have incorporated topical and emotional features to detect topical influencers [44] and social recommenders [132]. These methods account for a user's emotional alignment with a topic and their ability to influence peer group preferences, a particularly salient factor in politically charged or value-laden discussions. Network-based investigations of real-world political platforms further illustrate how influence manifests in practice. Studies analyzing Twitter, YouTube, and TikTok data (e.g., [109], [41], [1]) consistently find that a small subset of actors disproportionately shape discourse, often acting as opinion leaders, structural hole spanners, or community anchors. Importantly, their role is not static; it is shaped by platform-specific features such as retweet mechanics, algorithmic curation, and follower dynamics. Other works (e.g., Wang et al. [123]) emphasize the role of influencers in bridging ideological divides or exacerbating them by reinforcing echo chamber effects. Influencer detection thus becomes not just a question of visibility but of positionality within polarized spaces.

These findings converge on the view that accurate influencer identification must go beyond traditional centrality, incorporating contextual relevance, emotional resonance, and community topology. Particularly in

polarized networks, the most influential actors are often those who can bridge, rather than dominate, discourse communities.

2.2.3 Impacts of Influencers and Political Leaders on Community Opinion

The presence of influencers and political leaders on online platforms has significant consequences for the formation, polarization, and mobilization of public opinion. As social media becomes a primary arena for political discourse, influencers increasingly serve not only as information disseminators but also as opinion shapers, leveraging their reach and perceived authenticity to affect community sentiment and behavior. A growing number of empirical studies highlight how influencers, particularly those embedded in ideologically aligned communities, contribute to the reinforcement of partisan beliefs and the amplification of polarized narratives. For instance, Flamino et al. [41] analyzed Twitter discourse during U.S. elections and revealed that influencers often serve as focal points of political homophily, reinforcing filter bubbles through selective amplification of content. Similarly, Wang et al. [123] show that community leaders play distinct roles in echo chamber dynamics: while the former intensify intra-group consensus, the latter can bridge ideological divides but are less prevalent and less amplified by algorithms.

The content and strategies used by influencers/leaders also impact public opinion shaping. Riedl et al. [101] and Schmuck et al. [105] explored how influencers frame political narratives in ways that simplify complex issues, increasing political interest among youth but also contributing to superficial engagement or cynicism. This aligns with the findings of [68], who examined populist rhetoric on social media and demonstrated how leaders' emotionally charged, anti-immigrant narratives elicit strong affective responses, contributing to identity-based polarization. Simulation-based studies further validate these effects. For example, Betts et al. [15] demonstrated that influencer interventions can either accelerate or dampen polarization, depending on their alignment and dissemination strategies. Similarly, Helfmann et al. [62] used agent-

Table 2: Summary of Representative Approaches, Limitations, and Thesis Positioning

Category	Approaches	Strengths	Limitations	Thesis Positioning
Echo chamber detection	Network modularity, homophily, embeddings	Identifies polarized structures	Static; limited behavioral/temporal modeling	Dynamic behavioral indicators for echo-chamber evolution
Algorithmic radicalization	Recommender audits, diversity/exposure metrics	Reveals exposure bias	Mostly diagnostic; limited mitigation	Diversity diagnostics + learned graph rewiring (DRLGR)
Toxic behavior detection	Keyword and content-only classifiers	Scalable abuse detection	Weak context and emotion modeling	Sentiment and emotion-aware toxicity modeling
Influencer analysis	Centrality-based identification	Finds high-visibility actors	Visibility does not imply opinion influence	Sentiment-aware Friedkin–Johnsen opinion dynamics
This thesis	Behavioral, algorithmic, and influence models	Unified framework	–	Integrates structural, behavioral, and cognitive dimensions

based modeling to examine the effects of different media and influencer tactics, finding that even a small set of high-impact influencers can shift the opinion equilibrium of a community under the right conditions.

The target audience also influences the degree of impact. Research into youth political socialization (e.g., [86]) illustrates how influencers can shape not only opinions but also political behavior and civic engagement. However, these effects are not uniformly positive: some studies report increased ideological rigidity and lower tolerance for dissent among followers exposed to consistent one-sided messaging. Moreover, influence is often reciprocally constructed: audiences shape influencers just as much as influencers shape audiences. This dynamic is especially evident in platform-mediated environments like TikTok or YouTube [118]; [40], where algorithmic feedback loops incentivize the creation of emotionally resonant or extreme content, thereby entrenching existing divisions.

Collectively, the literature highlights that the influence of influencers and political leaders extends beyond mere dissemination of information.

Through their content strategies, network positions, and alignment with audience values, they fundamentally shape how opinions form, polarize, and persist, often transforming public discourse in unpredictable ways.

Chapter 3

Echo Chambers Detection

The proliferation of radicalized communities on online platforms has raised significant concerns due to their ability to reinforce extreme views, normalize harmful behaviors, and foster hostility against targeted groups. A key characteristic of such communities is the presence of two particularly critical manifestations: the formation of echo chambers and the spread of cyberbullying behaviors. This chapter addresses the challenge of identifying and analyzing toxic behaviors within radicalized communities through two complementary works.

The first work dedicated to detecting the formation and evolution of echo chambers, dynamic online environments where users are increasingly exposed to homogeneous content and reinforcing viewpoints. Understanding how these communities emerge and transform is critical, as echo chambers can exacerbate polarization, hinder exposure to diverse perspectives, and distort public discourse over time. Despite growing concern, prior research has focused mainly on static detection of echo chambers, emphasizing the structural properties of social networks or content similarity at fixed points in time. These approaches often overlook the temporal and behavioral dimensions of user participation, failing to capture how users' preferences and interactions evolve as echo chambers develop. This study addresses these limitations by asking two key questions: How can user behavior serve as a gauge to monitor the

evolution of echo chambers over time? And how well can behavioral patterns forecast changes in users' future community affiliations? To answer these, a multi-phase pipeline that integrates network embeddings, clustering techniques, and novel behavioral indicators was developed. This framework not only detects existing echo chamber structures but also predicts user migration between communities, offering a dynamic and behavior-informed perspective that advances beyond prior static or content-centric models.

The second work addresses the growing problem of toxic language, particularly cyberbullying within echo chambers, which often serve as incubators for radicalization and hostility. These tightly-knit environments can amplify aggressive behaviors by reinforcing shared grievances and normalizing harmful discourse. Cyberbullying encompasses a wide range of malicious behaviors, including harassment, defamation, exclusion, and threats [38], and its detection poses significant challenges. A key difficulty lies in the indirect, ambiguous, or context-dependent nature of abusive language, which often escapes traditional keyword- or rule-based detection models [129]. While prior work has focused mainly on surface-level textual features, such methods usually fail to capture the deeper emotional and psychological cues embedded in toxic interactions. To address these limitations, this study proposes a novel framework that integrates sentiment and emotion analysis alongside conventional textual signals. By examining emotional markers such as anger in gender-based attacks, disgust in ethnicity-related abuse, and fear in religion-based harassment, the model uncovers distinctive affective patterns that signal targeted hostility. This emotion-informed approach enhances the detection of toxic behavior in polarized environments, offering a more context-aware and fine-grained tool for understanding online aggression within echo chambers.

By addressing both structural toxicity—through the formation and evolution of echo chambers—and behavioral toxicity—through the detection of cyberbullying, this chapter provides an integrated perspective on the dynamics that sustain radicalized online communities. Together, these contributions offer a unified analytical framework that captures

the interplay between user interactions, language use, and emotional expression. This multidimensional approach advances our understanding of how toxic environments emerge and persist, shedding light on the mechanisms by which online spaces become increasingly polarized and hostile.

3.1 Echo Chamber Detection and Prediction in Radicalized Communities

Echo chambers have emerged as dynamic environments where users are exposed to homogeneous content as these communities evolve, with users shifting their behaviors and modifying their content preferences. Detecting echo chambers requires advanced analyses of user behaviors, including user language similarity, user interactions, and group relationships [70]. Building on the chapter's focus on structural toxicity, this section presents a framework for detecting and predicting the evolution of echo chambers within radicalized online communities. By analyzing user interactions, content preferences, and linguistic behaviors over time, the proposed pipeline identifies cohesive clusters of users with ideologically aligned views. It anticipates shifts in community membership based on their behavioral patterns. The approach combines network embeddings, clustering, and novel behavioral indicators such as user similarity, community closeness, and content alignment to provide both descriptive insights and predictive power regarding community dynamics.

3.1.1 Methodology

The proposed methodology includes multiple phases to analyze user interactions, content preferences, and language patterns across time, thereby enabling both descriptive analysis and predictive modeling of echo chamber dynamics. Figure 1 illustrates the proposed pipeline, which shows five main components.

Figure 1: Echo Chamber Detection Pipeline

A. Network Construction & Node Embedding

The pipeline starts with the graph construction phase. For each time frame i^{th}, a three-interaction network $G_i = (V_i, E_i)$ was built, where V_i represents the users, and E_i corresponds to replies made by users to others' comments. Once the graph is constructed, *GraphSAGE* is used to generate node embeddings for each user. This embedding process takes two inputs: A represents the graph adjacency matrix and V represents the embeddings of user comments, which are derived using BERT [32]. These inputs are fed through multiple *GraphSAGE* layers, combined with ReLU activation functions and batch normalization (see Figure 2). The result is a set of embedding vectors that capture user behavior for each time frame.

B. Clustering

After generating the embedding vector for each user, the *k*-means algorithm is used to partition users into groups. To validate the clustering efficiency, both the elbow method and the silhouette were employed by

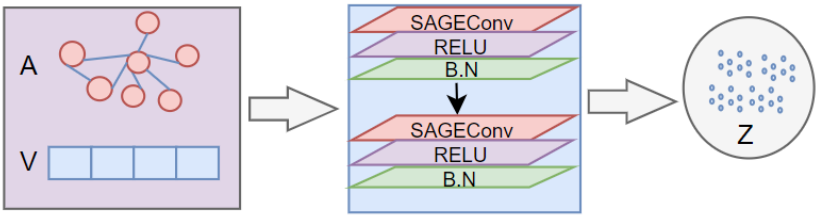

Figure 2: GraphSAGE Architecture

determining the optimal cluster number (K) for each i^{th} and validating the quality of the clustering.

C. Echo Chamber Indicators

Once the clusters were obtained for each time frame, the next phase focused on assessing the presence of echo chambers and tracking their evolution over time. To achieve this, a set of indicators has been designed to capture different behavioral aspects. These include the similarity between members, the position of users within the community, user behavior, and the alignment between users' comments and the associated videos.

Indicator 1: User Similarity (Sim). This indicator captures the linguistic behavior of users within their communities. It aims to quantify the degree of similarity in language among members of the same cluster; in other words, it reflects the level of linguistic homogeneity within a community. To measure this, cosine similarity is used to compute the average similarity between users' language patterns within each cluster. User comments were first embedded using BERT, and pairwise cosine similarities were then calculated between each user and all others in the same community. A high average similarity suggests a shared language style among users, which may indicate the presence of an echo chamber. AvgSim governed by Equation (3.1):

$$\text{AvgSim}_C = \frac{1}{N_C} \sum_{i=1}^{N_C} \text{Sim}_i \tag{3.1}$$

Here, $N_C$is the number of users, and Sim_i represents the average cosine similarity between user u_i and all other users in the same cluster.

Indicator 2: User Closeness (UCL). The user closeness indicator is designed to measure the distance between individual users and the ideological center of their community. To achieve this, the distance between each user's embedding vector and the centroid of their assigned cluster is calculated, where the centroid is assumed to represent the most ideologically extreme point within that cluster. This approach enables us to

determine whether users are moving closer to or further away from the dominant ideological position of their group. UCL governed by Equation (3.2):

$$UCL_i = \text{dist}(\text{Centroid}_{C_1}, \text{U}_i) - \text{dist}(\text{Centroid}_{C_2}, \text{U}_i) \tag{3.2}$$

Here, dist$(\cdot, \cdot)$ is the Euclidean distance, U_i is the user's embedding, Centroid_{C_1} and Centroid_{C_2} are the community centroids. A UCL_i closer to -1 indicates stronger alignment with the user's own group, suggesting deeper ideological conformity.

Indicator 3: Video Consumption Score (VCS). VCS indicator estimates user content preference, based on the assumption that commenting reflects content consumption. It is calculated as:

$$VCS_i = 1 - \frac{Total_{Ct_1}}{\text{Total\_Comments}} \tag{3.3}$$

Here, $Total_{Ct_1}$ is the number of comments a user posted on videos from category Ct_1, and Total_Comments is the total number of comments they made. VCS =0 indicates exclusive interaction with Ct_1, and 1 reflects full engagement with other categories.

Indicator 4: Comment-Video Alignment Score (CVS). CVS indicator measures how closely a user's comment aligns with the content of the video they watched. Cosine similarity is used for this purpose. A score near 1 indicates strong alignment, 0 reflects no alignment, and -1 implies opposing views. CVS governed by Equation (3.4):

$$\text{CVS}_i = \text{C.S}(C_i, V_j) \tag{3.4}$$

where C_i represent the user comments and V_j the video description.

D. User behavior prediction

After computing the proposed indicators, these indicators were used as input features to train a model that predicts users' future decisions, specifically whether they will remain in or leave their communities. A Deep Neural Network (DNN) was developed, consisting of six layers

with the following number of neurons: 1024, 256, 128, 32, 16, and 1, respectively.

3.1.2 Experimentation

This section presents a series of empirical investigations aimed at validating the proposed metrics designed to detect echo chambers and evaluating their effectiveness in forecasting shifts within online communities. Specifically, it outlines two key experiments and the associated data preparation procedures. The initial experiment investigates the capability of the indicators to identify and characterize echo chamber states. The second experiment examines patterns in user engagement with video content over time and assesses the predictive performance of these indicators in anticipating changes in user behavior.

A. Dataset Description & Preparation

A dataset initially compiled by [99] was employed, which contains a collection of videos categorized according to distinct political orientations. This study focuses explicitly on right-wing user communities, particularly those engaging with content labeled under the Alt-lite and Alt-right classifications—categories often associated with pronounced echo chamber effects. To facilitate temporal analysis, the dataset was segmented into three distinct periods: 2016, 2017, and 2018, based on the timestamps of user comments. Only users who were active and commented consistently across all three years were retained for analysis. This filtering process yielded a refined subset of 2,786 unique users who collectively posted 671,030 comments on a total of 9,001 videos. An overview of the dataset used in the experiments is presented in Table 11.

Category	Value
Videos no	9,001
Comments no	671,030
Unique users no	2,786

Table 3: Dataset Summary

To perform the analysis, three separate networks were constructed for each time frame. The structure of these networks is outlined as follows: users change their behavior. The 2016 network comprised 2,786 users and 2,200 connections; the 2017 network retained the same number of users but had 3,843 connections; and the 2018 network maintained the user count, with an increase to 4,904 connections. Subsequently, the *GraphSAGE* algorithm was employed to compute node embeddings. The process began with pre-processing user comments, which involved cleaning and tokenization, followed by embedding using a BERT-based model to capture nuanced semantic representations. In parallel, an adjacency matrix was constructed to represent user interaction patterns within the network. These semantic and structural features were then combined and used as input for the *GraphSAGE* model. Once the node embeddings were generated, the K-means clustering algorithm was applied, resulting in two distinct user groups for each temporal snapshot. The assumption was that each cluster corresponded to the same community across all three time periods. The clustering outcomes revealed a shift in community composition: for Cluster 0, the number of users declined from 1,885 to 1,530 in 2018, while it grew from 901 to 1,256 users during the same interval for Cluster 1. This pattern suggests simultaneous processes of user attrition and growth within the respective clusters.

B. Indicators Correlation

To validate the proposed indicators, their correlation with a benchmark metric, known as the Echo Chamber Score (ECS) [76], is examined. ECS quantifies the degree of community cohesion and separation by measuring the spatial distances between users in the embedding space. This metric was computed independently for each of the three temporal segments. The resulting ECS values illustrate the increase of echo chambers for the clusters, following an upward trend—from 0.8003 in 2016, rising to 0.8114 in 2017, and reaching 0.8417 in 2018—indicating an increasing polarization and insularity within user communities. Next, to examine the alignment between the proposed indicators and the ECS, the Pearson correlation coefficient was used to measure the statistical relationship be-

tween them. Each indicator is reviewed individually, beginning with an analysis of its distribution across network users, followed by its correlation with the ECS values.

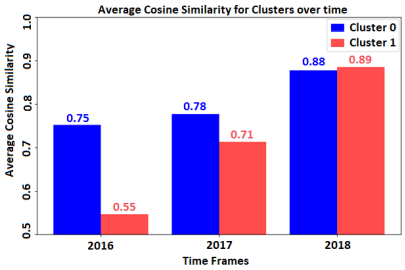

Figure 3: User similarities results

User Similarity. Mean similarity values within each cluster were computed using Equation (3.1). As shown in Figure 3, both clusters exhibit a steady rise in intra-cluster similarity across the three-year time frames. Notably, Cluster 1 demonstrates a more pronounced increase compared to Cluster 0, indicating a stronger trend toward linguistic alignment among its members. To understand the correlation between the Sim indicator and echo chamber formation, Pearson correlation coefficients were used as shown in Table 4. The Sim metric shows a robust positive correlation with the ECS in both clusters. This observation highlights the dynamic nature of echo chamber development, suggesting that shifts in users' language patterns are closely tied to changes in community cohesion and polarization over time.

User Closeness. The closeness score, calculated according to Equation 3.2, reflects how tightly users are positioned around their respective cluster centroids. Figure 4 illustrates the temporal evolution of this metric for Clusters 0 and 1. Cluster 0 is characterized by a broader spread of user positions, indicating greater internal variability. Despite this dispersion, the distribution remains centered within the interval $[-0.8, -0.7]$ across all time frames. In contrast, Cluster 1 exhibits a consistent upward trend in closeness scores, indicating a growing concentration of users near the cluster's center, which is indicative of increasing internal cohesion. The

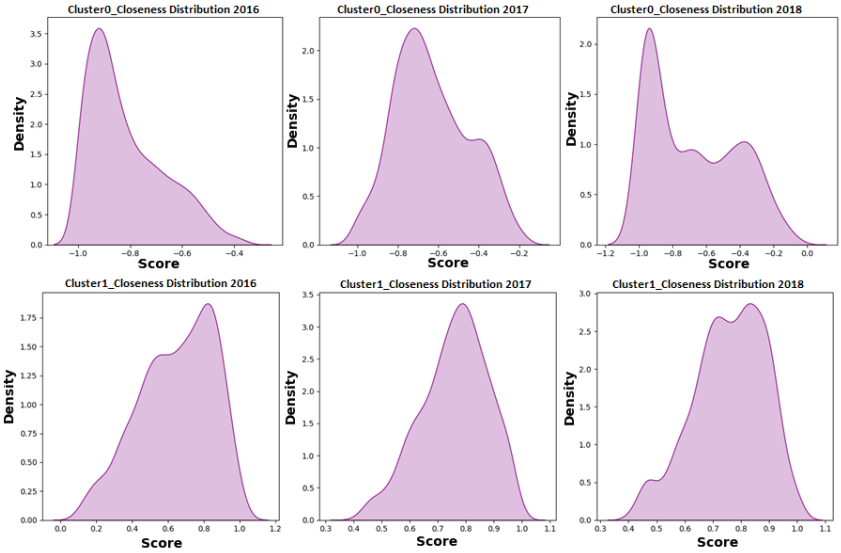

Figure 4: UCL distribution for each cluster

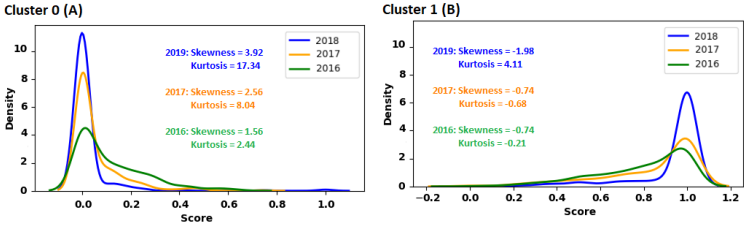

Figure 5: VCS values for both clusters.

	ECS	
	Cluster 0	Cluster 1
avgSim	0.95	0.97
avgUCL	0.88	0.93

Table 4: ECS correlation with avgSim and avgUCL.

correlation analysis between UCL and ECS, also summarized in Table 4, reveals a strong positive association for both clusters. This pattern highlights the significance of spatial user alignment in assessing echo chambers. Specifically, the data suggest that as users converge ideologically toward the center of their respective communities, the intensity of the echo chamber effect tends to increase.

VCS. To quantify user engagement with video content, the Video Consumption Score (VCS) was computed using Equation (3.3). The distribution of this metric across users in the network was assessed using descriptive shape statistics, specifically skewness and kurtosis, as outlined in [17, 79, 21]. Skewness measures the asymmetry of a distribution, while kurtosis reflects the degree of peakedness, with higher values indicating a greater concentration of data around the mean. As illustrated in Figure 5, both clusters exhibit increasingly peaked VCS distributions from 2016 to 2018, suggesting a rising concentration of users around central consumption behaviors. Additionally, skewness values indicate growing asymmetry in 2018 compared to 2016, with users diverging more noticeably toward both ends of the distribution spectrum.

The correlations between VCS metrics and the Echo Chamber Score (ECS), shown in Table 5, demonstrate a strong positive relationship between kurtosis and ECS for both clusters. Interestingly, while Cluster 0 exhibits a notable positive correlation with skewness, Cluster 1 shows a significant negative correlation with skewness. Elevated kurtosis is indicative of more centralized consumption patterns, reinforcing the presence of cohesive echo chambers. Conversely, skewness—whether positive or negative—implies a preferential leaning toward particular content types, further contributing to the intensity of the echo chamber. Overall, higher kurtosis and extreme skewness values are consistently associated with stronger characteristics of an echo chamber.

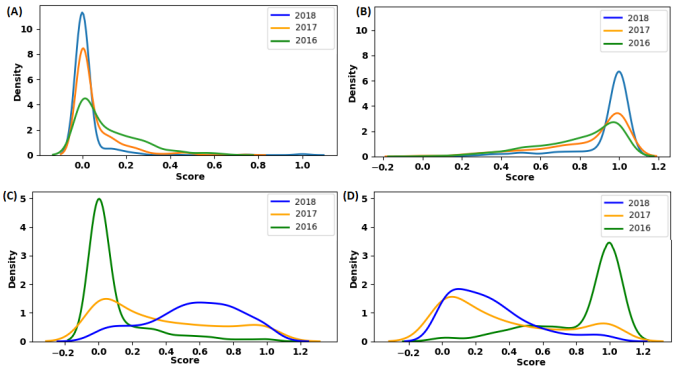

Figure 6: The users that stayed in cluster 0 and cluster 1 are presented in (A) and (B), while users moved from 0 to cluster 1 and vice versa are represented in (C) and (D), respectively.

CVS. Equation (3.4) was utilized to calculate the Community Variation Score (CVS). Analogous to the VCS indicator, statistical shape measures were applied to analyze the distribution of CVS across both clusters. The results indicated a largely random distribution pattern over time, revealing no significant trends that could effectively describe changes within the user communities. As reflected in Table 5, which presents the correlation between CVS and the ECS, no relationship between CVS's kurtosis and skewness values and the ECS for both clusters. This suggests that CVS exerts minimal impact on assessing echo chamber dynamics at the community scale.

From these observations, it is evident that the proposed indicators vary in their predictive power. Specifically, User Similarity, Closeness, and VCS show strong and consistent correlations with ECS within both clusters, highlighting their relevance in capturing community-level echo chamber dynamics. In contrast, CVS exhibits only a weak association with ECS, indicating its limited utility in evaluating echo chamber status at the collective level. Consequently, the first three indicators are more dependable for monitoring community evolution within echo chambers. Conversely, CVS may be better suited for analysis at the individual user level, offering nuanced insights into personal behaviors and preferences,

	ECS			
	Kurtosis		Skewness	
	Cluster 0	Cluster 1	Cluster 0	Cluster 1
VCS	0.87	0.89	0.84	-0.93
CVS	-0.17	-0.24	-0.03	-0.25

Table 5: ECS Correlation with VCS and CVS

such as video consumption patterns and content biases within the communities to which users belong.

C. Predicting community membership changes

This phase begins with the premise that users tend to reassess their community affiliations over time, reflecting shifts in preferences or opinions. Before evaluating the predictive capacity of the proposed indicators, the distribution of the Video Consumption Score (VCS) across users within communities was analyzed over three distinct time periods. The VCS is designed to capture individual consumption patterns and preferences longitudinally. Following this analysis, user membership predictions were conducted to assess the extent to which the indicators can effectively identify user affiliation decisions.

Video consumption analysis. We study the relationship between VCS and community membership for two user groups: those who remain consistent members of their communities and those who transition between communities. Understanding this relationship is critical because accurate predictions of community membership imply a corresponding ability to forecast changes in viewing behavior. To investigate this, we use the VCS (see Equation 3.3) to extract detailed insights into content interaction within these groups. Initially, users were divided into two categories: the first group consisted of individuals who remained in the same community continuously from 2016 to 2018. The second group consists of users who were part of one cluster in the first time frame and became part of another in the last time frame.

The VCS trends for users who remained in their original communities are presented in Figure 6[A-B]. It is evident that their consumption patterns strongly align with the defining content characteristics of their commu-

nities, as previously identified at the community level. Notably, the distribution for Cluster 0 narrows over time, indicating an increasing concentration of users who prefer videos rated between 0 and 0.1. Cluster 1 exhibits a comparable pattern, converging towards a score of 1, indicative of a growing affinity for alt-right content. These trends suggest that, over time, users are increasingly consuming information closely tied to the ideological leanings of their communities.

For users who changed community affiliation, their VCS scores are shown in Figure 6[C-D]. Users transitioning from Cluster 0 in 2016 to Cluster 1 in 2018 initially consumed only alt-lite content in the first time frame. The analysis reveals that, during the second time frame, users expanded their content consumption to include alt-right material. By the third time frame, their engagement with alt-right content had further increased. Conversely, users transitioning from Cluster 1 to Cluster 0 initially focused on alt-right content in 2016 but shifted their attention predominantly to alt-lite videos by 2018. These observed transitions reflect significant shifts in content preferences, offering valuable insight into how users' media consumption evolves alongside changes in their community memberships. This behavior illustrates the fluid and dynamic nature of user preferences, which adapt considerably as individuals migrate between ideological clusters.

User behavior prediction evaluation. To predict the shifts in community membership, the DNN model performance was examined across three distinct input setups: (1) user comments only, (2) a combination of comments and the proposed indicators, and (3) indicators alone. In this experiment, the dataset was annotated so that each sample was categorized as "stay" if the user was a member of the same community in all time frames, and move " if the user transitioned to a different community between the first and last time intervals. The data were partitioned into three sets: training, validation, and testing sets. Table 6 presents the model's performance metrics across these input scenarios. The model achieved an accuracy of approximately 86% when using only the indicators as input features, outperforming the 63% accuracy achieved when relying solely on user comments and the 68% accuracy achieved when

	Comment	Comment + Indicators	Indicators
Accuracy	0.63	0.68	0.86
F1_Score	0.58	0.64	0.85
Recall	0.63	0.68	0.86
Precision	0.59	0.67	0.86

Table 6: DNN Performance

combining comments and indicators. These findings demonstrate that incorporating the proposed indicators significantly enhances the model's capability to predict shifts in community membership.

3.2 Enhancing Cyberbullying Detection with Sentiment and Emotion Analysis

Cyberbullying is a prevalent form of toxic behavior within radicalized echo chambers. Effective cyberbullying detection requires not only advanced algorithms capable of interpreting nuanced language but also a comprehensive understanding of the emotional context behind interactions. This section introduces a deep learning–based approach that enriches traditional text features with sentiment and emotion analysis, enabling more accurate detection of abusive content. Initially, Sentiment and emotion analysis are integrated with conventional textual features to enhance the detection of cyberbullying content across various categories. In addition, a comprehensive investigation is conducted to examine the relationship between sentiment and emotional factors and specific types of cyberbullying, including those based on age, gender, ethnicity, and religion. The findings indicate that incorporating sentiment and emotion features enhances model performance, with a DNN trained on TF-IDF, sentiment, and emotion features yielding the most accurate results. Furthermore, the analysis reveals distinct emotional patterns associated with different forms of cyberbullying: age-based bullying is frequently linked with surprise, ethnicity-based bullying with disgust, gender-based bullying with anger, and religion-based bullying with both

fear and anger. Conversely, non-cyberbullying content generally exhibits a more neutral sentiment and emotional tone. This work contributes as follows:

- **Enhanced Detection with Sentiment and Emotion Analysis:** The detection process is enhanced by incorporating sentiment and emotion features into the models, with results demonstrating that these features contribute significantly to improved performance.
- **Insights into Sentiment and Emotion Patterns in Cyberbullying:** The relationship between sentiment, emotion, and specific types of cyberbullying was explored, revealing distinctive emotional and sentiment patterns for different bullying categories.

3.2.1 Methodology

This section introduces the approach employed, which encompasses multiple phases, as illustrated in Figure 7: Data Preprocessing, Feature Extraction, and Classification Modeling.

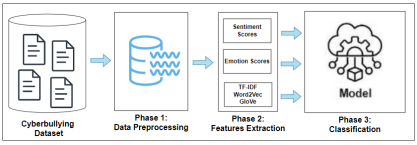

Figure 7: Cyberbullying detection methodology.

Data Preprocessing

The first component of the proposed pipeline is data preprocessing, which ensures that the data is clean and well-structured, thereby optimizing model performance. The preprocessing steps are outlined as follows:

- Handling Missing Data: Entries with missing data were removed to prevent potential bias in the model. This ensures that all training instances contain complete information.

- Language Filtering: Non-English words were removed from the dataset to focus solely on English text.

- Label Encoding: Categorical features were transformed into a numerical format using one-hot encoding, a necessary step for detection algorithms that require numerical input data.

- Text Cleaning: This step is used to improve data quality for the feature extraction process. The following techniques were applied:

  Eliminating Punctuation and Special Characters: Punctuation, symbols, and special characters were stripped out to ensure that only the meaningful textual content was analyzed.

  Eliminating Stop Words: Frequently occurring words that offer little value to the text's overall meaning (such as "and," "the," and "is") were removed to minimize data noise.

  The text was segmented into individual tokens.

  Lemmatization was performed to reduce words to their base form.

Feature Extraction

Following the preprocessing phase, the next step is feature extraction, which involves extracting a numerical representation from the text. Effective feature extraction is essential for capturing the key characteristics of the text, enabling the model to differentiate between various classes accurately. To identify the most effective representation, three feature extraction techniques—TF-IDF, Word2Vec, and GloVe were applied and compared. Additionally, emotion and sentiment scores were computed for each entry to provide supplementary features. Each method is described below:

TF-IDF: TF-IDF is a statistical measure used to assess the importance of words in a document relative to the entire corpus. It works by combining the term frequency, which quantifies how often a word appears in a specific document, with the inverse document frequency, which reduces the

weight of commonly occurring words across the corpus, thereby highlighting terms that are both frequent and distinctive.

Word2Vec: It is a neural network technique that generates dense vector representations, or embeddings, for words by considering their context. It captures semantic relationships between words by placing similar words in proximity within the vector space.

GloVe: GloVe is another word embedding technique, which, unlike Word2Vec, focuses on global word co-occurrences rather than local word co-occurrences, deriving word vectors from the global co-occurrence matrix of a corpus. GloVe combines the strengths of matrix factorization methods and local context-based learning. The embeddings are obtained from the ratios of co-occurrence probabilities, which indicate how frequently two words appear together in the same context throughout the entire corpus. GloVe generates dense vector representations similar to Word2Vec, but its focus on global co-occurrence statistics enables it to capture more comprehensive semantic relationships between words.

Emotion and Sentiment Scores: In addition to traditional feature extraction techniques, emotion and sentiment scores were generated for each text to enrich the feature set. These scores were derived as follows:

- Emotion Scores: The DistilRoBERTa-base model [61] was used to generate emotion scores, identifying various emotions in each tweet. DistilRoBERTa is a distilled version of RoBERTa [75], optimized for speed and accuracy, and is effective at extracting nuanced emotional states from text.

- Sentiment Scores: The RoBERTa model was employed to provide sentiment scores that classify tweets as positive, negative, or neutral. Twitter-RoBERTa is well-suited for capturing the informal and succinct nature of tweets, effectively identifying sentiment nuances in social media language. .

By enriching the feature sets with sentiment and emotion scores, the models gained the ability to capture the underlying emotional and attitudinal nuances present in the text more effectively. This augmentation

enables classifiers to distinguish between subtle variations in tone and intent.

Classification Modeling

The final phase of the proposed approach involves selecting and applying suitable classification algorithms for detecting cyberbullying. Several prominent algorithms were evaluated, including XGBoost, KNN, RF, LR, NB, SVM, and DNN. The model's performance was assessed using the standard metric, the F1-score, to select the algorithm that provides the best balance of performance for cyberbullying detection.

3.2.2 Experimentation

Experimental Setup

Dataset: A publicly available dataset sourced from Kaggle [87] was used to perform detection experiments. This dataset was extracted from Twitter and comprises over 47,000 tweets collected during the COVID-19 pandemic. It is labeled into several cyberbullying classes: gender, age, religion, ethnicity, and non-cyberbullying. The dataset is balanced across all categories, ensuring that the model does not favor any particular class, thus promoting fair training and evaluation.
Models Implementation: Each of the ML/DL models discussed in the Classification Modeling subsection was implemented using various combinations of feature extraction techniques. The goal was to identify the most effective combination for detecting cyberbullying. The implementation process was structured into three primary experiments, each comprising multiple sub-experiments. Each classification algorithm was executed with various combinations of feature extraction methods (TF-IDF, Word2Vec, GloVe) alongside supplementary features including emotion and sentiment scores, as detailed below:

Experiment 1: In the first experiment, GloVe embeddings were utilized as the primary feature extraction method. Each classification was evaluated using the following inputs: GloVe features, GloVe + Emotion,

GloVe + Sentiment, and GloVe + Emotion + Sentiment.

Experiment 2: The second experiment focused on TF-IDF as the feature extraction method. Each classification model was evaluated using TF-IDF features, TF-IDF + Emotion, TF-IDF + Sentiment, and TF-IDF + Emotion + Sentiment.

Experiment 3: In the third experiment, Word2Vec embeddings were used as the primary feature extraction technique. Each classification model was run with the following inputs: Word2Vec features, Word2Vec + Emotion, Word2Vec + Sentiment, and Word2Vec + Emotion + Sentiment.
Hardware and Software Environment: All experiments were conducted using Google Colaboratory (Colab), a cloud-based computational platform provided by Google. The execution environment was equipped with an Intel Xeon CPU, approximately 12 GB of RAM, and an NVIDIA GPU (Tesla T4) when required for deep learning experiments. The operating system was a Linux-based environment managed by Colab. All models were implemented using Python 3, with machine learning and deep learning libraries including Scikit-learn, TensorFlow/Keras, and XGBoost. Natural language processing and feature extraction tasks were carried out using standard Python libraries such as NLTK and Gensim.

Experimental Results:

Table 7 summarizes the findings of the first experiment, revealing that the DNN and XGB models consistently outperformed other models across all inputs. The highest F1 score was achieved by DNN when combining GloVe with both emotion and sentiment scores, reaching a value of 0.7865. Similarly, XGBoost also performed strongly, with its accuracy peaking at 0.7845 under the same combination. Both DNN and XGBoost showed significant gains when emotion and sentiment features were added to GloVe, demonstrating the effectiveness of incorporating these affective features into the model.
The outcome of Experiment 2, as illustrated in Table 8, shows that replacing GloVe with word2vec embeddings results in slight improvements in

Table 7: Model Performance with GloVe Feature inputs

Model	GloVe	GloVe + Emotions	GloVe + Sentiment	GloVe + Emotions + Sentiment
SVM	0.6458	0.6571	0.6584	0.6826
XGB	0.7212	0.7528	0.7785	0.7845
KNN	0.6088	0.6282	0.6276	0.6368
RF	0.6829	0.7157	0.7239	0.7237
LR	0.6937	0.7169	0.7726	0.7378
NB	0.3609	0.3957	0.4030	0.4281
DNN	**0.7122**	**0.7713**	**0.7850**	**0.7865**

the performance of most models. This once again highlights the superior performance of the DNN and XGBoost models. The DNN model achieved an accuracy of 0.8021 with word2vec alone, which increased to 0.8216 when word2vec was combined with emotion and sentiment scores. XGBoost also demonstrated significant performance, reaching an accuracy of 0.7988 with the complete feature set. The addition of emotion and sentiment scores notably enhanced the model's performance, especially for DNN and XGBoost. However, other models, such as Random Forest and Logistic Regression, also experienced moderate improvements. Naive Bayes remained the least effective model, showing minimal gains across different inputs.

Table 8: Model Performance with word2vec Feature inputs

Model	word2vec	word2vec + Emotions	word2vec + Sentiment	word2vec + Emotions + Sentiment
SVM	0.5475	0.5884	0.6075	0.6555
XGB	0.7755	0.7799	0.7895	0.7988
KNN	0.6808	0.6870	0.6894	0.6923
RF	0.7329	0.7463	0.7429	0.7494
LR	0.6977	0.7097	0.7139	0.7188
NB	0.2572	0.2584	0.2618	0.2630
DNN	**0.8021**	**0.8139**	**0.8155**	**0.8216**

For the third experiment, Table 9 reports the results of using TF-IDF as the primary feature extraction method, along with emotion and sentiment scores. In this context, employing TF-IDF led to noticeable improvements in model performance. Table 9 demonstrates that integrating TF-IDF with emotion and sentiment features in a DNN model achieves the highest accuracy. Using the same feature set, the RF model attained a maximum accuracy of 0.9677, while XGBoost obtained an accuracy of 0.9314.

Table 9: Model Performance with Different Feature inputs

Model	TF-IDF	TF-IDF+ Emotions	TF-IDF+ Sentiment	TF-IDF+ Emotions + Sentiment
SVM	0.7153	0.7269	0.7215	0.7854
XGB	0.9015	0.9125	0.9166	0.9314
KNN	0.8248	0.8524	0.8617	0.8904
RF	0.9258	0.9394	0.9365	0.9677
LR	0.8845	0.8954	0.8988	0.9148
NB	0.8355	0.8765	0.8758	0.8754
DNN	**0.93**	**0.9592**	**0.9547**	**0.9853**

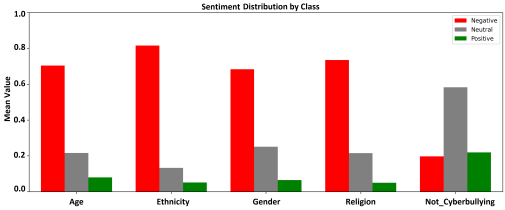

Figure 8: Distribution of sentiments across cyberbullying categories.

The above experiments demonstrate that the combination of TF-IDF with emotion and sentiment scores, particularly when used with the DNN model, yields the most accurate and reliable outcomes for cyberbullying detection. The DNN model consistently outperformed others across different feature sets, with its peak performance achieved in Experiment 3.

Table 10: Comparison of Performance with Various Approaches

Reference	Proposed Appraoch	Performance
[77]	Random Forest + TF-IDF	0.8245
[7]	Random Forest + GloVe	0.9354
[117]	RoBERTa + GLOVE + PCA features	0.9877
[88]	Decision Tree + BoW + PSO	0.9654
[124]	SOSNet + SBERT	0.9277
[103]	BERT	0.9734
Proposed	**TF-IDF+ Emotions + Sentiment**	**0.9890**

Comparison With Previous Studies: After concluding that models incorporating TF-IDF with emotion and sentiment scores as input performed the best (as demonstrated in Experiment 3). Table 10 presents the results of a comparison with other models, highlighting that the proposed method achieves superior performance with an F1 score of 0.9890,

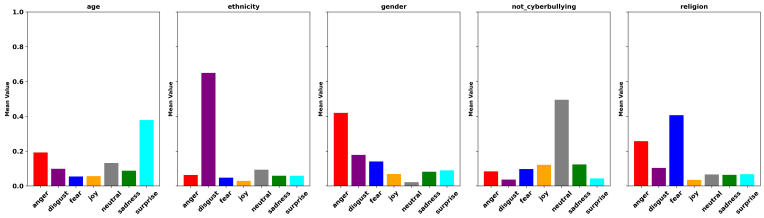

Figure 9: motions associated with cyberbullying and non-cyberbullying categories

surpassing several notable approaches in the field. For instance, [77] reported 0.8245 using Random Forest + TF-IDF, while [7] achieved 0.9354 with Random Forest + GloVe. Even advanced methods, such as RoBERTa with GloVe and PCA proposed by [117], which achieved a score of 0.9877, were slightly outperformed by the model. Other methods, such as [88] (Decision Tree + BoW + PSO) and [103] (using BERT), also performed well but did not reach the level of accuracy achieved by the proposed approach. This comparison underscores the effectiveness of integrating sentiment and emotion analysis for enhanced cyberbullying detection.

Further Analysis

This section provides an in-depth analysis of both sentiment scores and emotion scores, aiming to highlight whether particular emotional states or sentiment tendencies are more strongly associated with specific types of cyberbullying (e.g., age, ethnicity, gender, religion, and general non-cyberbullying content).

Figure 8 presents the sentiment distributions. The highest mean values of negative sentiment are across all cyberbullying categories. In contrast, neutral sentiment is most frequent in the 'Not Cyberbullying' class, demonstrating that general, non-bullying content is more balanced in tone. Positive sentiment is relatively scarce across all categories, further emphasizing the overwhelming presence of harmful content within the cyberbullying landscape.

The analysis of emotion scores, as illustrated in Figure 9, reveals the

dominant emotional states associated with each category of cyberbullying. As depicted, the surprise emerges as the predominant emotion in the "age" class, while disgust leads in both "ethnicity" and "religion" classes. Anger is prominently associated with "gender" cyberbullying. Across all categories, negative emotions dominate, with positive emotions, such as joy, registering very low values, especially in comparison to negative emotions like fear and sadness, which exhibit varying levels of influence depending on the class. For the "Not Cyberbullying" class, the emotional distribution is more balanced, with neutral and joy having relatively higher values compared to the cyberbullying categories. However, negative emotions such as anger and sadness persist. Still, their influence is significantly lower, suggesting that non-bullying content evokes a broader range of emotional responses, with a more substantial presence of positive and neutral emotions compared to the highly negative emotional landscape in cyberbullying cases. This pattern of emotional distribution suggests that the different forms of cyberbullying are more frequently characterized by intense, negative emotional states, reinforcing the harmful nature of these online behaviors.

3.3 Discussion

The findings of this chapter highlight the dynamic nature of echo chambers in radicalized online communities, providing empirical evidence that user behaviour, specifically linguistic similarity, community closeness, and content consumption, plays a central role in their evolution. By integrating behavioral indicators with network embedding techniques, the proposed framework not only detects echo chamber structures but also predicts user migration patterns across communities. This extends the existing literature, which has focused mainly on static structural or content-based detection of echo chambers, demonstrating the added value of behavioral proxies in capturing temporal and interactional dynamics.

These results refine our understanding of echo chamber formation by highlighting the importance of affective and linguistic alignment in sus-

taining ideological homogeneity. Whereas prior work has emphasized network topology and selective exposure as drivers of polarization, this thesis shows that linguistic convergence within groups is strongly correlated with measures of community cohesion. The strong predictive performance of behavioral indicators also suggests that echo chambers can be conceptualized not only as structural phenomena but as evolving socio-linguistic ecosystems shaped by both user behavior and algorithmic mediation.

Practically, the proposed framework offers actionable insights for platform governance and moderation. The ability to forecast whether users are likely to remain in or exit echo chambers enables the development of early-warning systems to flag communities at risk of radicalization. Moreover, the integration of sentiment and emotion analysis for detecting toxic behaviors such as cyberbullying highlights the potential of combining behavioral and affective cues to improve moderation efforts in hostile environments. These findings have implications for cybersecurity practitioners and policymakers seeking proactive approaches to mitigate online polarization and extremist recruitment.

A limitation of the empirical analysis presented in this chapter is the temporal scope of the dataset, which spans the period 2016–2018. Since then, social media platforms have undergone substantial changes in recommendation algorithms, moderation policies, and content formats, including the rise of short-form video and intensified algorithmic curation. The specific quantitative patterns observed in the experiments may not fully reflect contemporary platform dynamics in 2025. However, the proposed methodology is grounded in user behavioral, linguistic, and interactional signals rather than platform-specific algorithmic implementations. These signals represent stable mechanisms of social reinforcement and polarization that are expected to persist across platform evolutions. Nevertheless, caution is warranted when generalizing empirical magnitudes, and future work should validate the proposed framework on more recent datasets and emerging platforms to assess its robustness under evolving socio-technical conditions.

As a result, this chapter directly addresses RQ1 by showing that echo

chambers can be effectively detected and their evolution forecast using behavioral indicators, thereby advancing the thesis's first contribution to monitoring and understanding radicalized communities.

Chapter 4

Algorithmic Radicalization and Mitigation

Social media platforms are increasingly relying on algorithmic curation to optimize user experiences. RS now plays a central role in shaping the content users consume, directly influencing their preferences, opinions, and behaviors [96, 66]. While personalization can enhance engagement and satisfaction, it also carries significant risks: reinforcing filter bubbles, amplifying biased narratives, and guiding users down narrow and potentially radicalized pathways—particularly when diversity in recommendations is low or skewed [94, 102, 12]. Despite increasing awareness of these risks, a substantial gap remains in understanding the underlying mechanisms by which recommendation systems contribute to radicalization. Prior studies have primarily addressed this issue through ML and DL approaches, typically framing radicalization as a content moderation or misinformation problem. However, such treatments often overlook the deeper structural and behavioral dynamics that govern how radical content is surfaced, reinforced, and consumed. As a result, current solutions may fail to account for the algorithmic amplification of ideology over time. This gap has motivated a growing body of research aimed at safeguarding informational diversity and preserving democratic resilience in digital spaces [67, 26, 126, 14, 122, 72].

This chapter introduces a dual-layered framework that addresses algorithmic radicalization as both a diagnostic and intervention problem. The first layer investigates the systemic pathways through which personalization mechanisms amplify ideological homogeneity. It examines how recommendation algorithms interact with user histories and content features to generate self-reinforcing trajectories. To capture this, a diversity-aware analytical framework was proposed, which integrates explainable AI and predictive modeling to trace and analyze recommendation trajectories. A central contribution of this layer is the development of the Recommendation Diversity Score, a novel metric designed as an early warning indicator of ideological enclosure. By quantifying the decline in content diversity along recommendation paths, the diversity score is shown to correlate with escalating radicalization risk. The analysis demonstrates that reduced diversity in recommendations, especially when combined with high engagement, serves as a predictive signal for future ideological extremity.

The second layer operationalizes proactive, context-aware interventions. We introduce DRLGR, a deep reinforcement learning-based model that dynamically rewires recommendation graphs to mitigate radical exposure. Unlike heuristic or rule-based approaches, DRLGR learns adaptive policies that optimize content plurality and actively reduce the likelihood of radicalization. It triggers intervention once a user's radicalization trajectory surpasses a defined risk threshold. Extensive experiments on YouTube and news recommendation datasets validate the effectiveness of this approach. DRLGR consistently outperforms existing baselines in both reducing radical content exposure and sustaining long-term diversity in user content streams. These findings underscore the potential for adaptive, data-driven interventions to reshape recommendation dynamics in safer and more equitable ways.

Together, these two contributions offer an integrated framework for understanding and addressing algorithmic radicalization. By diagnosing how personalization mechanisms reinforce ideological extremity and introducing a scalable mitigation strategy through reinforcement learning, this chapter provides actionable insights into the development of safer,

more transparent, and socially responsible recommendation systems.

4.1 Analyzing Radicalization Pathways in YouTube Recommendations

This contribution addresses a critical gap in understanding the radicalization pathway within RS by presenting a framework for diagnosing radicalization pathways within YouTube's recommendation network. Two primary research questions guide the study:

- RQ2.1. Which primary content features predominantly influence the selection of the next video from recommended lists?
- RQ2.2. To what extent do the attributes of viewed content and users' watching histories contribute to forecasting the level of diversity in video recommendations?

To address these research questions, a multi-stage framework was introduced, powered by explainable artificial intelligence (xAI), to analyze recommendation trajectories using YouTube as a case study. The first phase involves constructing a video recommendation graph, where graph-based embedding techniques are employed to capture both relational and semantic properties of the content. In the second phase, explainable AI methods, specifically LIME, are utilized to identify the most influential video attributes (e.g., political bias, topical category, and textual metadata) that contribute to the propagation of radicalized recommendations. Building upon these insights, the third phase introduces a composite diversity indicator, operationalized as the range of topics, viewpoints, and content characteristics presented to users. A predictive DL model is then developed to estimate diversity levels within recommendation lists, incorporating both content features and dynamic user behavior patterns. This model integrates user watch histories with the previously extracted video attributes, taking into account temporal shifts in user interests. Empirical evaluation demonstrates the model's efficacy

in capturing diversity dynamics and highlights its potential utility in signaling early-stage radicalization pathways. This approach provides a robust method for detecting early signs of algorithmic radicalization and advances the interpretability of RS through feature attribution.

4.1.1 Methodology

This section introduces a three-stage methodological pipeline designed to uncover and quantify how YouTube RS contributes to the radicalization path through patterns of content diversity. The proposed methodology is presetned in in Figure 10, it encompass: (1) the creation of a recommendation network along with the extraction of graph-based embeddings for videos, (2) an explainable link prediction approach to determine the significance of different features in forecasting connections between videos and their role in promoting radicalized recommendations; and (3) the prediction of recommendation diversity by utilizing the feature importance scores from the second stage to evaluate how these factors affect the range of recommended videos.

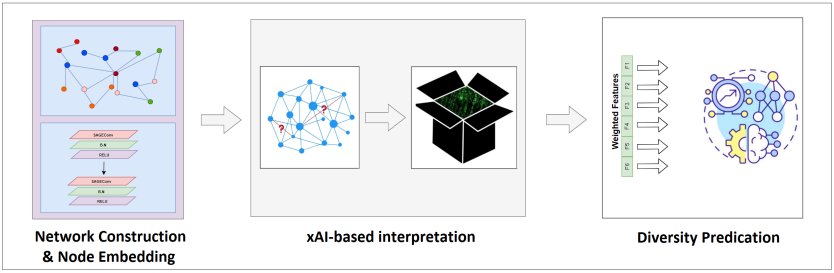

Figure 10: The architecture of the proposed three-stage methodology.

Network Construction & Node Embedding

The proposed pipeline begins with constructing the network by forming a directed network $G = (V, E)$ that captures video-to-video recommendation links, where V dis the videos, and E comprises edges that signify

the recommendation connections between these videos. Each node in the network is associated with its video metadata.
Following the construction of the recommendation graph, the GraphSAGE framework [56] was used to derive low-dimensional embeddings for nodes that effectively represent the graph's structural properties. GraphSAGE creates these embeddings by aggregating information from each node's neighbors. The GraphSAGE framework captures the local topology within a two-hop radius around every node, thereby extracting pertinent local features essential for forecasting the next recommended video. The entire graph is fed into the model simultaneously, producing embeddings that integrate both the graph's connectivity patterns and node-specific characteristics. These resulting embeddings serve as supplementary video features for further processing steps.

xAI-based Interpretation

In the second stage, the mechanisms underlying radicalized recommendations are examined using a two-step approach: (1) predictive modeling of radical recommendation links using link prediction, and (2) feature importance interpretation through explainable AI (xAI) techniques. Utilizing explainable AI (xAI) techniques to determine feature significance offers a deeper understanding of how video attributes impact these predictions and contribute to the formation of radicalized recommendation trajectories.
Link Prediction The network edges were categorized into two types: *radicalized* and *non-radicalized* instances. The edge is labeled as *radicalized* when both connected entities share the same ideological affiliation using Equation (4.1). Conversely, if the connected entities come from differing affiliations, the connection is marked as *non-radicalized* as specified by Equation (4.2). The feature set for each edge is formed by merging the characteristics of its source and target nodes with their associated node embeddings. These combined features serve as input to a Random Forest classifier, which is tasked with predicting the likelihood of edges existing between node pairs.

$$E_p^+ = \{(u, v) : u, v \in V, u \neq v, e_{u,v} \in E, \mathrm{C}(u) = \mathrm{C}(v)\} \tag{4.1}$$

$$E_n^- = \{(u, v) : u, v \in V, u \neq v, e_{u,v} \notin E, \mathrm{C}(u) = \mathrm{C}(v)\} \tag{4.2}$$

where E_p^+ and E_n^- are the *radicalized* and *non-radicalized*, respectively, , $\mathrm{C}(u)$ and $\mathrm{C}(v)$ correspond to the category of nodes u and v, respectively.

xAI Interpretations To gain insight into the internal reasoning of the classifier and evaluate the contributions of various video features, the LIME [100] method was applied. LIME approximates the model's local behavior by perturbing input features and observing the resulting changes in predictions. LIME is leveraged to generate feature attributions for a set of test samples and compute average importance scores across all predicted links. This procedure produces an interpretable ranking of the most influential features driving radical recommendation predictions, thereby increasing transparency on how user watch histories and video characteristics shape the algorithmic recommendation pathways.

Diversity within Recommendations

The final step in the approach involves defining and forecasting a metric that quantifies the extent of radicalization in the recommendations. The *Diversity Score* (D) is presented and subsequently assessed for its validity. This stage is pivotal as it emphasizes diversity as a fundamental measure for evaluating radicalization within video RS. Specifically, a low diversity score indicates that users are interacting with a limited range of content. In contrast, a high diversity score reflects a wide variety of recommended content. The motivation behind this metric is to provide a systematic means to evaluate the diversity inherent in the recommended content.

Diversity Score. In this context, diversity represents the breadth of content presented to users in alignment with their interests, aiming to measure the variety of viewpoints encountered within their recommended videos. This score is calculated across the entire set of recommendations by considering several video characteristics, such as the political bias score (indicating each video's ideological leaning), the political bias category the topic category of the video (categorizing content by subject matter), and the embedding of the video description (which encodes the textual information of the description). These features are part of the dataset and elaborated upon in the Dataset Section. Formally, it calculated using Equation (4.3):

$$\begin{aligned} D = \frac{1}{5}\big(&D_{\text{political bias score}} + D_{\text{political bias category}} \\ &+ D_{\text{video topic category}} \\ &+ D_{\text{video description embedding}} \\ &+ D_{\text{node embedding}}\big) \end{aligned} \tag{4.3}$$

First, the diversity for each attribute across all videos within the list was computed. These individual attribute diversity values are then combined to form a comprehensive diversity metric for the entire recommendation set, as expressed in Equation (4.3). By equally weighting each attribute, the method ensures that various factors contribute fairly to the overall assessment of diversity. The diversity is computed for each specific attribute as follows:

- $D_{\textbf{political bias score}}$**:** The diversity in political bias is calculated using the variance formula shown in Equation (4.4). Variance measures the spread of values, with higher variance indicating greater diversity and lower variance indicating more homogeneity in political bias scores.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{4.4}$$

Where n is the number of videos in the recommendation list, X_i represents the political bias score of the i^{th} video, and $\bar{X}$ is the mean political bias score across all videos in the list.

- $D_{\textbf{political bias category}}$ **and** $D_{\textbf{video topic category}}$**:** For these categorical attributes, diversity is computed using the information entropy formula in Equation (4.5). Higher entropy values correspond to higher diversity, reflecting a more even distribution across categories.

$$E = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{4.5}$$

Where n is the number of distinct categories, and p_i is the proportion of videos in the recommendation list that belong to the i^{th} category.

- $D_{\textbf{video description embedding}}$ **and** $D_{\textbf{node embedding}}$**:** Diversity for these continuous vector attributes is based on dissimilarity, measured via cosine similarity. First, the average pairwise cosine similarity among all embeddings in the recommendation list is computed as shown in Equation (4.6). Then, diversity is defined as one minus this average similarity, thereby capturing overall dissimilarity.

$$\begin{aligned} \text{AvgSim}(k) &= \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \text{Sim}(i, j) \\ D(k) &= 1 - \text{AvgSim}(k) \end{aligned} \tag{4.6}$$

Where $\text{Sim}(i, j)$ denotes the cosine similarity between the embedding vectors of videos i and j, and N is the number of videos in the recommendation list.

After calculating the diversity scores for each attribute, the overall diversity of the recommendation list is computed using Equation (4.3), which is a simple average of these individual scores.

Indicator Validation. After introducing the Diversity Score, its relationship with the level of radicalization is assessed. To this end, a Radicalization Score is established to quantify the dominance of videos from specific ideological groups within a recommendation list. A recommendation list is classified as "radicalized" when the majority of its videos belong to one specific ideological category. Formally, the Radicalization Score (RS) for a given category is calculated using Equation (4.7):

$$RS_c = \frac{\text{count of videos in category } c}{\text{total count of videos in the recommendation list}} \tag{4.7}$$

where c denotes the ideological category.
Next, the ORS is defined by Equation (4.8) as the difference between the maximum and minimum category-specific radicalization scores:

$$ORS = \max_c(RS_c) - \min_c(RS_c) \tag{4.8}$$

The ORS indicates which recommendations skew toward extreme viewpoints. A score of 0 indicates no radicalization (i.e., uniform distribution across categories), whereas a score close to 1 indicates intense radicalization concentrated in particular ideological groups.
Predicting Diversity within Recommendations After confirming the validity of the diversity score, the next step focuses on forecasting this score for the recommendation lists presented to individual users. To achieve this, a deep neural network (DNN) is utilized, which takes multiple inputs as shown in Figure 11. The model incorporates six inputs. These features are combined with importance weights obtained from explainable AI (xAI) techniques, capturing subtle aspects of radicalization as outlined in the methodology, alongside user viewing histories. A detailed description of these features follows:

- **Political Bias Score of Video:** A quantitative value indicating the video's ideological stance.
- **Political Affiliation Category of Video:** A categorical label denoting the political grouping assigned to the video.

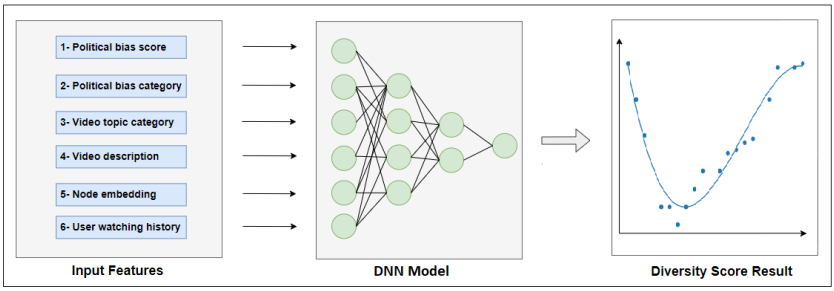

Figure 11: Architecture of the predictive model.

- **Topic Category of Video:** The subject area under which the video is classified.
- **Video Description Embedding:** A vectorized representation capturing the semantic content of the video's textual description.
- **Node Embedding of Video:** A learned vector embedding representing the video's position within the recommendation graph structure.

The above features are redefined as an xAI-weighted feature using the Equation (4.6):

$$\mathrm{F}_n = F_i \times W_i^P \tag{4.9}$$

where F_n is the weighted (new) feature value, F_i is the original feature value, and W_i^P is the LIME-derived feature weight.

- **Watching History Score (WHS):** This feature quantifies user watch behavior The score is derived from the political bias scores of the videos the user has watched, which range from -1 to 1. To compute WHS, an exponential decay weighting scheme is applied to assign higher weights to more recently watched videos, ensuring the latest watched video has the highest influence, which is defined as:

$$y(t) = y_0 \cdot e^{-kt} \tag{4.10}$$

where $y(t)$ is the weight at time t, y_0 is the initial weight at $t = 0$, k is the decay constant controlling the decay rate, and $e \approx 2.71828$ is Euler's number.

Once weights are assigned, the WHS is computed using Equation (4.11):

$$\text{WHS} = \sum_{i=1}^{n} P_i \times W_i \tag{4.11}$$

where n is the total number of videos watched, P_i is the political bias score of the i^{th} video, and W_i is its corresponding weight.

These features serve as inputs to the DNN, enabling it to predict the diversity of the recommendation list.

4.1.2 Experimental Evaluation

The details of the experimental procedures undertaken to assess the method's capability in forecasting the diversity score associated with YouTube recommendation lists are presented in this section.

Experimental Setup

Dataset Overview and Processing A dataset containing videos and their corresponding recommendations from the YouTube platform [59] was employed, where each video is labeled according to its distinct political leaning. It includes data from 100,000 synthetic user profiles, each simulating specific ideological biases by consuming a set of 100 videos randomly selected to reflect their assigned political orientation. For evaluation, the recommendation paths were tracked by recording both the homepage and the subsequent "up-next" suggestions for each sock puppet, starting from an initial seed video selected from a predetermined list. This tracking continued for up to 20 consecutive recommendations,

with the primary analysis focusing on the 'up-next' video suggestions. In total, the dataset comprises 100,000 accounts, 23,735 unique videos watched, and over 300,000 videos recommended. Each video is annotated with various metadata attributes, including an ideological bias score, political category, topic classification, and textual description. The ideological bias score ranges between -1 and 1.
Using this ideological slant metric, videos are assigned to one of several political categories, including left, center-left, center, center-right, and right. Additionally, videos are categorized by topic into various Each video also includes a textual description that serves as metadata capturing the content details.

Table 11 summarizes the dataset used in the experiments.

Table 11: Dataset Description

	Training Phase	Testing Phase
Videos no	23,735	381,153
Channels no	1,256	111,450

For the experiments, a directed graph was constructed to model video-to-video recommendation relationships, containing 106,436 nodes and 826,315 edges. To enhance the graph representation, various video-related features were incorporated as node attributes. These features spanned multiple data types, including textual, numerical, and categorical variables. Categorical data were transformed into a numerical format, and a BERT model [33] was applied for textual attributes to extract meaningful embeddings. The preprocessing pipeline for the descriptions involved converting all text to lowercase for consistency, removing special characters, and then tokenizing the text. The cleaned text was then input into BERT to generate semantic-rich vector embeddings, which were subsequently assigned to their corresponding graph nodes. To further capture structural and contextual information, GraphSAGE was used to generate embedding vectors. These embeddings were integrated as supplementary video features utilized in the experimental evaluation outlined in later sections.

Explainable AI (xAI) Interpretation Framework. Following the creation of the graph and the extraction of node embeddings, an xAI analysis was conducted to indicate the impact of each feature on the formation of radicalized recommendation edges. This analysis was divided into two core phases: predicting links within the graph and interpreting feature contributions.

To perform link prediction, several ML models were evaluated. Among these, the RF model was selected to predict connections between nodes based on its performance metric. Following the link prediction step, the LIME algorithm was employed to interpret the RF model's outputs, highlighting the significance of various input features. A detailed summary of the prediction outcomes and corresponding LIME-based explanations is provided in the Results.

Diversity Prediction Methodology. The second experimental phase focused on predicting diversity scores for recommendation lists using a DNN. The DNN was trained and evaluated on three distinct input configurations: video-only attributes, user watch history exclusively, and a combination of both. The network architecture consisted of four layers with sizes 256, 32, 16, and 1 neuron(s), respectively. This configuration was selected based on empirical testing of multiple architectures. Diversity scores computed for individual user recommendation lists served as the target variable. Model optimization was performed using MSE as the loss, and predictive performance was quantified by the coefficient of determination (R^2).

Experimental Outcomes:

xAI-Based Interpretation Results

The results of the explainable AI (xAI) experiments, concentrating on the performance of link prediction and the resulting analysis of feature importance. As summarized in Table 12, the Random Forest classifier

outperformed other models, achieving an F1_score approaching 89%. Figure 12 presents the feature importance scores derived using LIME, revealing the relative influence of each feature on the predictions. Notably, the "Political Bias Score" is a highly influential factor, with an importance rating of nearly 0.9, underscoring its critical role in identifying radicalized recommendation links. Conversely, the "Political Category" feature exhibited considerably lower importance, suggesting a less prominent effect. Other features displayed comparable contribution significance.

The strong predictive performance of the link prediction model confirms its ability to effectively capture the relationships between videos, thereby supporting the reliable interpretation of feature effects through xAI techniques. These findings provide a valuable basis for the next phase of experimentation.

Table 12: Link prediction performance

Model	F1_score
SVM	0.65
KNN	0.71
LR	0.55
ANN	0.83
RF	**0.89**

Table 13: Correlation between proposed Diversity and Radicalization score

	Radicalization Score
Diversity Score	-0.84

Prediction Results

The validation outcomes for the proposed metric were presented, followed by the results from the prediction experiment. To assess the validity of the diversity indicators, a Pearson correlation coefficient was calculated between the two indicators. The results show a strong inverse

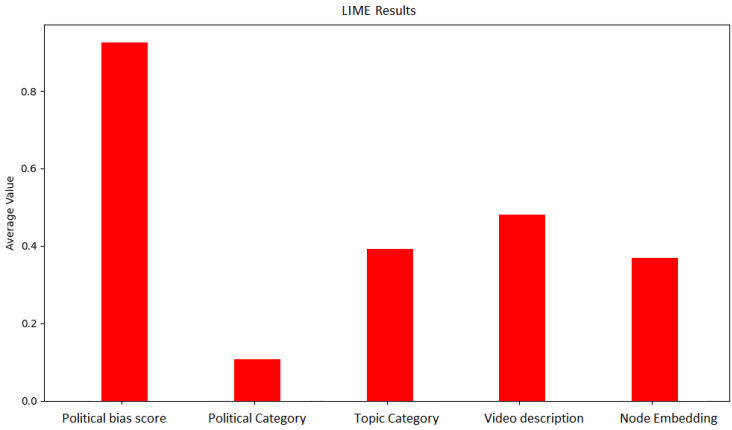

Figure 12: xAI LIME Results.

linear association in Table 13. This suggests that higher diversity scores are associated with proportionally lower degrees of radicalization. In essence, these findings demonstrate that the diversity metric serves as an effective proxy for measuring radicalization levels based on the established correlation.

Following the validation of the indicator, the predictive model's outcomes are presented in Table 14 and Figure 13. Figure 13 illustrates both training and validation loss curves. As detailed in Table 14, ANN achieved a 0.87 R^2 score when utilizing a combination of video attributes and user watching history (WHS) as inputs. In contrast, employing only video attributes yielded an R^2 score of 0.78, while using WHS alone produced a lower score of 0.61. These results highlight the enhanced predictive power gained by integrating both feature sets, emphasizing the importance of their interplay in influencing user content exposure. This insight is crucial for understanding the risk of RS.

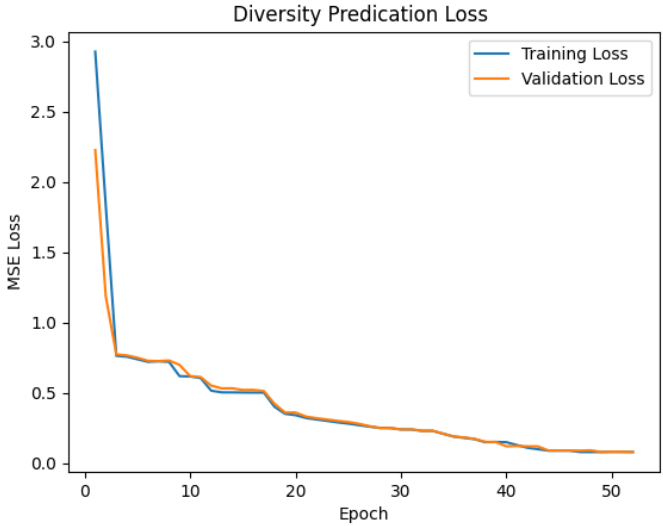

Figure 13: Prediction model performance.

Table 14: Model Performance

	Video Attributes	WHS	Video Attributes + WHS
R2 score	0.78	0.61	0.87

4.2 Mitigating Radicalization through Adaptive Graph Rewiring

Building on the diagnostic insights from analyzing radicalization pathways, this section shifts focus from detection to intervention. While understanding the mechanisms that drive radicalized content consumption is essential, it is equally critical to develop strategies that actively reduce users' exposure to extremist material while preserving the quality of recommendations.

Several conventional methods that focus on content moderation are proposed to include these risks, although these efforts are frequently unfea-

sible. In contrast, some researchers have focused on refining the RS to reduce the spread of radicalized content while maintaining the user experience [51, 37]. [39] proposed a greedy algorithm to reduce segregation within the RS through rewiring graph edges. The authors define segregation as the probability that users will encounter benign information after initially starting from a malicious node. Similarly, [27] investigated how an edge rewiring method in a directed graph may reduce exposure to radical content. They proposed a process called the GAMINE algorithm to address this issue.

While these two efforts [39] and [27] provided methods to reduce the harmful content in RS, they have significant drawbacks. First, measures for radicalization based on topology metrics, such as the number of steps in a random walk, have failed to cover the content within RS. Second, the rewiring process in their approaches is based on predefined criteria and not learned from the dynamics of RS. In this study, an intervention framework was employed that dynamically rewires connections involving radicalized users, redirecting them toward neutral content nodes.

In the methodology, the Radicalization Score $Rad(G)$ was introduced, a metric that measures the diversity within RS as a proxy of measuring the radicalization level [14], unlike the segregation score [14]. Additionally, to reduce it, the DRLGR method was proposed, which performs rewiring edges in the RS graph when the users are at high $Rad(G)$ risk, unlike previous heuristic approaches that rely on fixed rules. DRLGR employs RL to develop an adaptive edge rewiring policy tailored to the context. This policy is optimized to maximize the reduction of the $Rad(G)$ through continuous interaction with the environment, learning from feedback and rewards obtained from prior actions.

The experiments performed on different datasets to validate the effectiveness of the DRLGR approach: one representing the YouTube RS and another focused on news article recommendations. In each case, DRLGR was benchmarked against multiple baselines. The findings show that the proposed method reduces the $Rad(G)$ within the recommendation graph, achieving consistent improvements over time.

This approach reframes radicalization mitigation as a structural and

behavioral challenge, treating recommendation networks as dynamic systems rather than static datasets. It demonstrates that targeted, learning-driven rewiring can achieve sustained reductions in radicalization, highlighting the potential of anticipatory algorithmic governance to safeguard informational diversity on a large scale. This study contributes the following key advancements:

- Propose a metric for measuring the radicalization in RS
- Introduced a method to reduce radicalization effectively.

4.2.1 Methodology

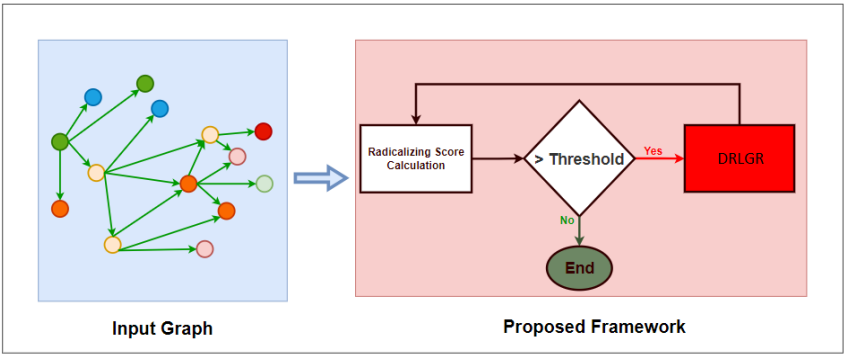

Figure 14: The Proposed Framework

The detailed methodology is described in this section. As shown in Figure 14, we first compute the Rad(G), and then the DRLGR technique is used to reduce the $Rad(G)$ if it exceeds a preset threshold. Until the graph's $Rad(G)$ drops below the designated threshold, the following procedure is done recursively.

- **Rad(G)**: Random walks are used to determine the graph's initial radicalization score.
- **DRLGR**: Rewiring the edges is done by activating the RL agent; otherwise, nothing has to be done.

- **Rad(G*)**: After each rewiring process, a new measurement for the radicalization score was computed for the redesigned graph G^*.

This process continues until the score becomes below the target level if G^* stays above the threshold.

Network Construction Preliminaries

Assume a network $G = (V, E)$ that is composed of several nodes V and several edges E. The nodes are divided into two parts: radicalized nodes, represented as R, and neutral nodes, represented as N, while E represents the recommendation links between V. Each node V in the network has a fixed number D; subsequently, the recommendation list of node v is denoted by $L(v)$. This graph is weighted, in which W_{ij} between v_i and v_j corresponds to the number of times node v_i is recommended by node v_j. For modeling the user behavior on G, the random walk is employed, in which it interacts with G through a path $\pi_u = \{v_1, v_2, \ldots, v_n\}$, where each $v_i \in V$ represents a node visited sequentially. Transitions between nodes follow a Markov chain defined by the transition matrix P. Each entry p_{ij} in P is the probability of moving from v_i to v_j. This first-order Markov assumption reflects the idea that a user's next choice depends primarily on the current content being viewed. The matrix P has dimensions $n \times n$.

$$P[i,j] = \begin{cases} p_{ij} & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \tag{4.13}$$

where

$$P_{ij} = \frac{W_{ij}}{\sum_{k \in V} W_{ik}},$$

and W_{ij} is the weight between v_i and v_j.

A notable vulnerability in such networks lies in the emergence of *radicalized* pathways, where a user becomes confined within a trajectory $\pi_u = \{v_1, v_2, \ldots, v_n\}$ composed predominantly of radicalized nodes. To address this issue, this study aims to reduce the $Rad(G)$ value associated with such routes by restructuring the network G. The proposed restructuring involves a sequence of edge-rewiring operations K, wherein the $L(v)$ is modified by substituting an existing connection between two radicalized vertices $\{R_i, R_j\}$ with a link joining a radicalized and a neutral node $\{R_i, N_j\}$. Executing this transformation produces an improved, optimized network denoted as G^*.

Radicalization Score Calculation

Here, the description provides details of the process of computing $Rad(G)$. As explained in Figure 15, it begins by computing the $Rad_{User}(\pi_u)$ for each user path using Equation 4.14, and it measures the radicalization score associated along the route for each user. The sequence π_u represents the random walk trajectory on the G, in which it starts from any radicalized node and terminates after reaching three neutral nodes.

$$\text{Rad}_{\text{User}}(\pi_u) = \frac{1}{|\pi_u|} \sum_{i \in \pi_u} \text{Rad}_i \tag{4.14}$$

where Rad_i defined in Equation 4.15.

$$\text{Rad}_i = \frac{L(R)}{|L|} \tag{4.15}$$

where $L(R)$ is the number of neighbors labeled as radicalized, and L is the total number of neighbors.
Once $Rad_{User}(\pi_u)$ is calculated for each user, the $Rad(G)$ is computed by averaging the $Rad_{User}(\pi_u)$ values over all user paths, as specified in Equation 4.16.

$$\text{Rad}(G) = \frac{1}{|U|} \sum_{u \in U} \text{Rad}_{\text{User}}(\pi_u) \tag{4.16}$$

where U is the number of users, and π_u is the path associated with user u. $Rad(G)$ ranges from 0 to 1, where 0 indicates complete neutrality, and 1 signifies a high level of radicalization.

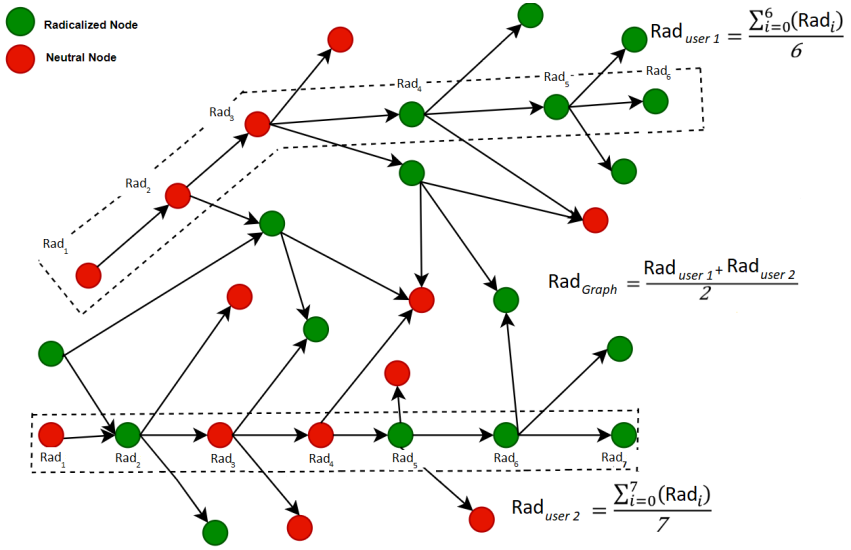

Figure 15: Radicalization score calculation

DRLGR Method

In the proposed framework, the DRLGR method serves as a fundamental element. It is based on an RL approach, utilizing a Deep Q-Network (DQN) to train the model to select edges for rewire, thereby maximizing the expected cumulative reward. DRLGR is activated at the terminal radicalized node when $Rad_{User}(\pi_u)$ surpasses a predefined threshold. It substitutes the edges between two radical nodes with an edge connecting the same source node to a neutral node. On the other hand, in the absence of such a candidate edge, the agent abstains from making changes during that iteration. Subsequently, the network's structure is updated and $Rad(G)$ is recalculated to evaluate the effectiveness of the intervention. DRLGR encompasses the following components:

Environment. The first component in the DRLGR is the environment, which represents the G that the agent interacts with to perform the actions. This environment is updated after each action performed. **Actions.** Each rewiring process performed by the agent represents an action in the DRLGR.
Rewards. After each rewiring action, the RL agent calculates a metric called the reward, which measures the effectiveness of the rewiring actions. It aims to maximize the reward:

$$\text{reward} = Rad(G) - Rad(G^*),$$

Policy. Here, the policy component represents the mechanism of choosing the rewiring action that maximizes the reward. The policy is learned through interaction with the environment.
Model Training. The DRLGR model is developed using the DQN algorithm [82] to perform the rewiring actions. By approximating the Q-value function, the agent can anticipate the expected cumulative reward of performing specific rewiring actions in various graph states. This capability guides the agent toward edge adjustments that effectively lower the radicalization metric $Rad(G)$ over time. During the learning phase, the agent decides on rewiring moves based on an epsilon-greedy approach, ensuring a balance between exploring new actions and leveraging known beneficial strategies. Each selected action alters the graph topology, followed by a calculation of rewards for the resulting variation in $Rad(G)$. These experiences are recorded in a replay memory, from which the agent randomly samples data to update the Q-network, enhancing the robustness and efficiency of the training process. Through repeated episodes, the agent progressively improves its edge-rewiring strategy by continuously learning from the consequences of its actions.

4.2.2 Experimentation

To assess the efficacy of the DRLGR framework in reducing $Rad(G)$ in RS, a series of experiments was carried out utilizing authentic datasets sourced from video-sharing websites and news media platforms. These

tests were structured to measure the effects of edge rewiring interventions initiated when individual users exceed predetermined thresholds for radicalization. Furthermore, the performance of DRLGR was benchmarked against standard baseline methods as well as a previously established heuristic algorithm.

Experimental Setup

Dataset Description & Preparation

The first dataset represents videos along with their corresponding recommendation links collected from YouTube [99]. Each video is labeled as radicalized or neutral, where radicalized videos are categorized into various political classifications such as "alt_right," "alt_lite," and "intellectual dark web". For experimental analysis, the dataset was split into two groups based on view counts: the "small" subset S containing videos with fewer than 100,000 views, and the "big" subset B containing videos with more than 10,000 views. From each subset, three recommendation networks were generated, varying the out-degree $D \in \{5, 10, 20\}$, representing the number of recommendations assigned to each video node. This allowed evaluation of the approach's sensitivity to different recommendation list sizes. The resulting graphs are named accordingly: for the small subset, YouTube-D5-S, YouTube-D10-S, and YouTube-D20-S; and for the large subset — YouTube-D5-B, YouTube-D10-B, and YouTube-D20-B The second dataset utilized in the experimental evaluation is the NELA-GT-2019 collection [89], which comprises English-language news articles aggregated from various publishers. Each entry in this dataset includes metadata such as the article's title, full text, URL, publication date, and source, and is annotated as either reliable or unreliable. For analysis, the dataset was segmented into three temporal subsets corresponding to August, September, and October. From each monthly subset, recommendation graphs were generated with a fixed out-degree of $D = 10$, resulting in the graphs NEWS-1 (August), NEWS-2 (September), and NEWS-3 (October). Table 15 summarizes the key characteristics of these graphs alongside the initial radicalization scores computed for both datasets.

Algorithms: The DRLGR algorithm was initially tested across a

Table 15: Constructed graph description for both datasets.

YouTube				
Subset Name	**D**	**Number of Nodes**	**Number of Edges**	**Initial Rad(G)**
YouTube-D5	5	29,732 100,536	119,112 363,604	0.83 0.84
YouTube-D10	10	29,732 100,536	238,244 727,208	0.88 0.87
YouTube-D20	20	29,732 100,536	476,448 1,454,416	0.91 0.89
NEWS				
Subset Name	**D**	**Number of Nodes**	**Number of Edges**	**Initial R(R0)**
NEWS	10	21,577 20,150 25,967	252,167 216,824 275,658	0.79 0.71 0.81

range of user radicalization thresholds $Rad_{User}(\pi_u)$ within the set $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ to examine its capability in minimizing the overall radicalization score $Rad(G)$. Following the evaluation at each threshold, the value that produced the most favorable outcomes was selected as the optimal cutoff point. The resulting model based on this threshold is denoted as $Model_1$. Thereafter, $Model_1$ was benchmarked against two baseline approaches, BSL_1 and BSL_2, in addition to a preexisting algorithm introduced in [39] to comprehensively assess its relative performance.

- The first baseline, BSL_1, maintains the RL framework utilized in $Model_1$ for edge rewiring decisions, but alters the graph input by replacing the original edge weights. Instead of using recommendation frequencies, BSL_1 calculates edge weights based on node similarity measures. Thus, this baseline serves as a variant of the DRLGR method applied to graphs weighted by similarity rather than interaction counts.

- The second baseline, BSL_2, is the DRLGR algorithm that performs the rewiring based on different criteria. Unlike the original ap-

proach, where only neutral nodes are chosen for rewiring, BSL_2 permits the RL agent to select nodes at random.

- Finally, the DRLGR approach is evaluated against the heuristic method HEU proposed by [39]. This comparison provides a comprehensive benchmark to gauge DRLGR's effectiveness relative to existing solutions.

Hardware and Software Environment: All experiments were conducted using Google Colaboratory (Colab), a cloud-based computational platform provided by Google. The execution environment was equipped with an Intel Xeon CPU, approximately 12 GB of RAM, and an NVIDIA GPU (Tesla T4) when required for deep learning experiments. The operating system was a Linux-based environment managed by Colab.

Experimental Results

DRLGR Performance Across Various Thresholds:
To investigate how different user radicalization thresholds affect the reduction of $Rad(G)$, the DRLGR algorithm was applied to both datasets using threshold values $Rad_{User} \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. The outcomes for several YouTube graph subsets at these thresholds are illustrated in Figure 16. The radicalization ratio is defined as

$$\frac{Rad_T}{Rad_0},$$

where Rad_0 is the initial radicalization score before any rewiring, and Rad_T is the score following K rewiring operations. The findings show that DRLGR effectively lowers the radicalization ratio as the number of rewiring steps K increases. Its most significant impact is observed at lower thresholds, specifically $Rad_{User}(\pi_u) = 0.5$ and 0.6. Conversely, at elevated thresholds such as $0.7, 0.8$, and 0.9, the reduction in radicalization is less pronounced, with the ratio remaining relatively high despite numerous rewiring actions.

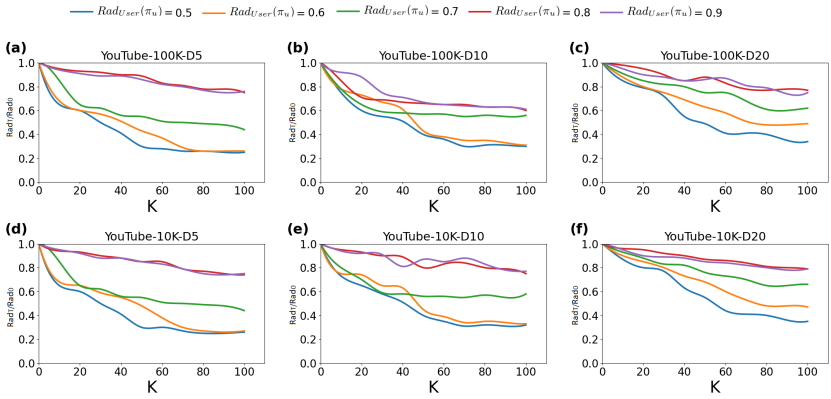

Figure 16: Performance of DRLGR with varying $Rad_{User}(\pi_u)$ threshold in first dataset.

Conversely, the results for the News Feeds dataset are depicted in Figure 17. Consistent with observations from the YouTube dataset, DRLGR achieves the most significant decrease in

$$\frac{Rad_T}{Rad_0}$$

at a user radicalization threshold of $Rad_{User}(\pi_u) = 0.5$ across all three monthly subsets. The efficiency of the model is further underscored by its ability to reach this reduction with roughly 25 rewiring steps. In contrast, a higher number of rewiring operations K is necessary at the 0.6 threshold. Moreover, the algorithm's capability to lower

$$\frac{Rad_T}{Rad_0}$$

diminishes considerably when applied at elevated thresholds of 0.8 and 0.9. These findings suggest that DRLGR performs optimally when activated at $Rad_{User}(\pi_u) = 0.5$, hereafter designated as $Model_1$ in the following sections.

Effectiveness of Proposed Algorithm: After establishing the effectiveness of $Model_1$ in lowering the $Rad(G)$, a comparison against BSL_1, BSL_2, and HEU was performed.

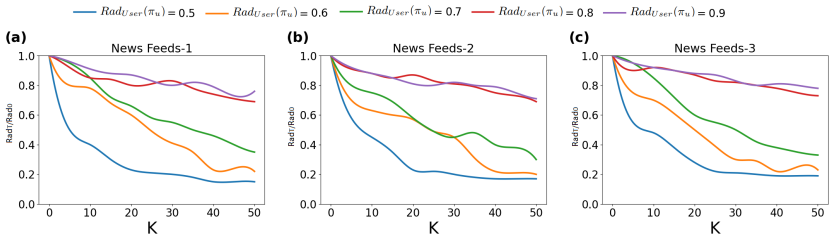

Figure 17: Performance of DRLGR with varying $Rad_{User}(\pi_u)$ threshold in second dataset.

To guarantee reliability, each experiment was conducted over five independent trials. Figure 18 displays the average results of $Model_1$ alongside HEU, BSL_1, and BSL_2 across variations of the graphs. The plotted curves represent mean performance over these runs, with shaded regions illustrating the standard deviation.

In comparisons with the baselines BSL_1 and BSL_2 on both subsets, $Model_1$ consistently outperforms BSL_2, highlighting the drawbacks of random rewiring. When set against BSL_1, which utilizes similarity-based edge weights, $Model_1$ demonstrates superior reductions approximately 70% after 50 rewiring operations for the $D = 5$ and $D = 10$ graphs. Although BSL_1 performs nearly as well, it remains slightly less effective than $Model_1$. For $D = 20$, $Model_1$ continues to maintain an edge over BSL_1, albeit requiring a larger number of rewiring steps.

A notable contrast emerges when comparing $Model_1$ with HEU in terms of long-term outcomes. Initially, HEU achieves more rapid reductions in radicalization, often surpassing $Model_1$ during the early phase (up to 25–30 rewiring actions), as evident in sub-figures (a) and (d). However, HEU's progress plateaus thereafter. In contrast, $Model_1$ sustains its downward trajectory, ultimately outperforming HEU in most scenarios. This advantage is especially pronounced in more intricate graphs, such as those in sub-figures (c) and (f), where $Model_1$ achieves roughly a 20% greater reduction in radicalization after 100 rewiring steps, demonstrating stronger resilience and superior long-term performance. Exceptions are visible in sub-figures (b) and (e), where HEU retains a persistent lead

over $Model_1$, suggesting that HEU excels in particular mid-sized graph configurations.

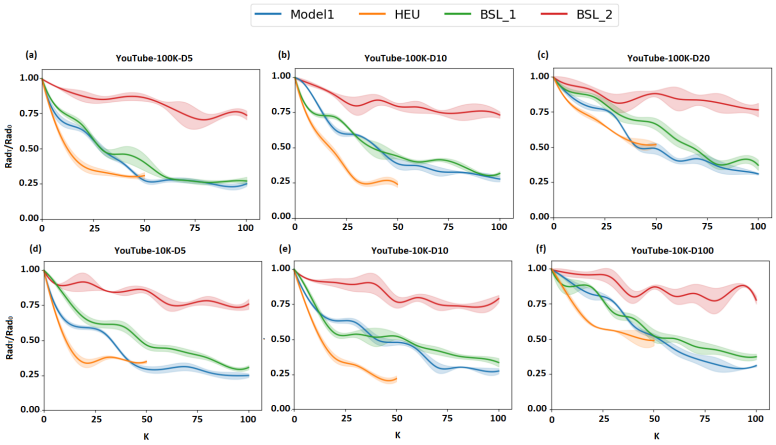

Figure 18: DRLGR vs BSL_1, BSL_2, and HEU in first dataset.

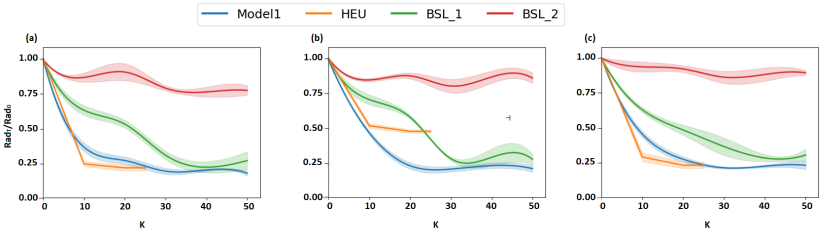

Figure 19: DRLGR vs BSL_1, BSL_2, and HEU in second dataset

Figure 19 illustrates the comparative performance of the proposed $Model_1$ against BSL_1, BSL_2, and HEU on the second dataset. Consistent with the observations from the YouTube dataset, $Model_1$ outperforms both baseline models BSL_1 and BSL_2 in all graph configurations. Specifically, $Model_1$ achieves a substantial reduction in the $Rad(G)$ of approximately 70% after 30 rewiring steps, while BSL_1 requires a greater

number of rewiring operations to reach comparable reductions. Meanwhile, BSL_2 remains largely ineffective, failing to significantly lower the $Rad(G)$.

For $Model_1$ and HEU, the latter demonstrates a strong initial impact, particularly in subfigures (a) and (c). Although HEU induces a rapid decline early on, $Model_1$ shows a steadier and more persistent improvement that ultimately surpasses HEU across all three scenarios. For instance, subfigure (a), which corresponds to the August dataset, reveals that HEU delivers a sharper reduction within the first 10–15 rewiring actions, but then its performance plateaus around 30 rewiring steps. In contrast, $Model_1$ achieves a final radicalization decrease of about 70%, exceeding HEU's 65%. In subfigure (b), representing September, $Model_1$ clearly outperforms HEU by attaining greater reduction with fewer rewiring steps. For the October data shown in subfigure (c), both models achieve similar levels of radicalization reduction. While HEU initially outpaces $Model_1$ in the early rewiring phase, $Model_1$ gradually closes the gap.

Overall, these results indicate that although HEU excels at quick early-stage intervention, $Model_1$ provides more durable and practical de-radicalization, particularly in larger, more complex recommendation networks.

Statistical Significance Analysis: To provide a rigorous assessment of the DRLGR method compared to the baselines HEU, BSL_1, and BSL_2 across both datasets, a statistical significance test was performed using t-tests at $K = 50$ and $K = 100$. Table 16 summarizes these results for both datasets. Against BSL_2, DRLGR exhibits consistently strong, with p-values well below 0.001 for every graph and rewiring step in the first dataset, and similarly robust significance levels ($p < 0.005$) across all NELA-GT graphs. When contrasted with BSL_1, DRLGR attains statistically significant enhancements at $K = 100$ across all YouTube graphs, while also showing substantial gains at $K = 50$ in several cases. The NELA-GT results follow a comparable pattern, where DRLGR significantly outperforms BSL_1 at $K = 50$ for all graphs. However, differences at earlier rewiring stages ($K = 25$) are minor or not significant, indicat-

Table 16: p-values result for DRLGR on both Datasets

YouTube Dataset				
Graph	**K**	**DRLGR vs HEU**	**DRLGR vs BSL_1**	**DRLGR vs BSL_2**
YouTube-D5-100K	50	0.360	0.143	**0.00017**
	100	–	**0.0028**	**0.00002**
YouTube-D5-10K	50	**0.0034**	0.866	**0.000019**
	100	–	**0.039**	**0.0000028**
YouTube-D10-100K	50	**0.00046**	**0.025**	**0.00069**
	100	–	**0.00043**	**0.0000063**
YouTube-D10-10K	50	0.055	**0.030**	**0.00031**
	100	–	0.056	**0.0000057**
YouTube-D20-100K	50	**0.039**	0.415	**0.00016**
	100	–	**0.028**	**0.000012**
YouTube-D20-10K	50	**0.0097**	**0.019**	**0.0000058**
	100	–	**0.026**	**0.0000075**
NELA-GT Dataset				
Graph	**K**	**DRLGR vs HEU**	**DRLGR vs BSL2**	**DRLGR vs BSL1**
NEWS-1	25	0.558	**0.000061**	0.066
	50	–	**0.00000042**	**0.036**
NEWS-2	25	**0.030**	**0.0047**	0.079
	50	–	**0.0000031**	**0.0016**
NEWS-3	25	0.703	**0.0014**	0.318
	50	–	**0.0000046**	**0.041**

ing that BSL_1 may offer some short-term competitiveness but does not maintain this advantage as rewiring continues.
The comparison with HEU reveals nuanced performance dynamics: on the YouTube dataset at $K = 50$, DRLGR achieves significant improvements in four out of six graph instances. However, for YouTube-D5-100K ($p = 0.360$) and YouTube-D10-10K ($p = 0.055$), the performance difference is not statistically significant, suggesting that HEU remains competitive during the initial stages of network rewiring. A comparable pattern emerges in the NELA-GT dataset at $K = 25$, where DRLGR significantly outperforms HEU only on NEWS-2 ($p = 0.030$), with no notable advantage observed on NEWS-1 and NEWS-3. These findings indicate that while HEU and BSL_1 can deliver comparable outcomes in the early phases, DRLGR consistently demonstrates superior results as the rewiring depth increases. The pronounced statistical significance at larger rewiring steps (e.g., $K = 100$) reinforces this conclusion, aligning with the patterns shown in Figures 18 and 19, and confirming that DRLGR's performance gains are both meaningful and robust rather than due to random variance.

4.3 Discussion

The first contribution introduces a principled diagnostic model that treats recommendation diversity as an early warning signal of radicalization. By integrating explainable AI with predictive modeling, it reveals latent dynamics that link user behavior and content variety, showing that the loss of diversity in recommendations is not accidental but predictive of ideological enclosure. Political bias emerged as the most influential attribute shaping radicalized recommendation pathways, surpassing factors such as topic category, video description, or node embeddings. This finding highlights political leaning within video content as a key driver of radicalization, emphasizing the need for monitoring politically biased material to curb the spread of radical content.

The analyses presented in this chapter confirm that recommendation systems can play a significant role in amplifying radicalization path-

ways, while also demonstrating that structural interventions can mitigate these risks. Specifically, the proposed diversity-based indicator effectively quantified users' exposure to radical content, revealing that algorithmic recommendations are not neutral but shaped by prior user behavior and video attributes. The reinforcement learning–based rewiring strategy (DRLGR) was demonstrated to enhance diversity within recommendation flows and decrease radicalization scores without compromising overall engagement.

These findings extend prior large-scale audits of YouTube's recommendation system, which have documented pathways toward extremist content. Unlike prior descriptive audits, however, this study contributes a computational framework that operationalizes radicalization risk as a measurable property of the recommendation graph. By coupling diversity assessment with adaptive rewiring, it advances beyond detection toward actionable mitigation. Theoretically, this aligns with recent scholarship on "algorithmic affordances", illustrating how small structural modifications in recommendation networks can produce macro-level shifts in ideological exposure.

Unexpectedly, the experiments revealed that user engagement levels remained stable even after increasing the diversity of recommendations. This challenges a common assumption in platform governance that interventions to counter radicalization necessarily reduce user satisfaction or activity. A plausible explanation is that users still encountered content relevant to their interests, but in a more ideologically varied format, suggesting that diversity and engagement need not be mutually exclusive. This finding holds promise for designing platform interventions that strike a balance between safety and business imperatives.

From a practical standpoint, the proposed framework offers a blueprint for platform designers and policymakers seeking to reduce harmful recommendation patterns. The results demonstrate that reinforcement learning–based strategies can dynamically adapt to evolving recommendation graphs, providing a scalable solution for mitigating radicalization pathways in real time. Such approaches could complement content-based moderation by targeting the structural conditions

that exacerbate the dissemination of harmful content.

It is important to note that these findings are based on simulations with synthetic agents, and engagement stability was observed in this controlled computational setting. While the DRLGR strategy maintained engagement mathematically, real-world users may respond differently due to cognitive dissonance or individual preferences. Consequently, these results should be interpreted as indicative of the potential effectiveness of structural interventions rather than definitive evidence of human behavior outcomes.

Altogether, this chapter answers RQ2 by demonstrating that recommendation systems contribute to online radicalization while also presenting a reinforcement learning–based mitigation strategy, thus fulfilling the thesis's second contribution to countering algorithmic amplification of extremism.

Chapter 5

Role of Influential Actors

This chapter shifts the focus to the underexamined driver of online radicalization: the human element. The mechanisms through which the persuasive influence of digital actors, including online influencers, political figures, and media personalities, operates within social networks. These individuals actively shape public discourse and accelerate ideological shifts through targeted strategies that exploit network structures, emotional appeals, and credibility signals. While prior research has explored influence in social networks from a structural perspective, such as network centrality or reach, these approaches often overlook critical dimensions of influence, including how it unfolds over time, how emotions interact with opinion shifts, and how persuasive language strategically shapes audiences. Consequently, understanding influence solely through static metrics risks underestimating the nuanced and multidimensional nature of digital persuasion, leaving both scientific inquiry and societal interventions incomplete.

This chapter addresses the overarching problem of how human-driven influence operates in social networks, integrating structural prominence, temporal dynamics, emotional resonance, and rhetorical strategy into a unified analytical framework. It tackles three interrelated subproblems:

(1) Measuring Influencer Impact Beyond Structural Centrality: Tradi-

tional centrality metrics (such as follower count, number of retweets, or static centrality scores) have long served as proxies for influence, failing to capture the actual capacity of individuals to shift collective opinions. This chapter introduces a dynamic, causally grounded approach to identifying influential users whose actions measurably alter the equilibrium opinions of their communities. The goal is to move beyond superficial measures of popularity and engagement, focusing instead on the real-world influence that shapes opinions and attitudes.

(2) Modeling Opinion and Emotional Dynamics Over Time: Influence is not static; opinions and emotions evolve as individuals interact. Existing models often overlook the temporal and emotional dimensions of social influence. This chapter develops a sentiment-aware opinion modeling framework, applying it to real-world social networks to examine how influencers shape both opinion trajectories and emotional climates over time. By comparing communities with and without key influencers, the study reveals how selective actors can amplify emotional volatility, polarize opinions, and shape community sentiment in subtle yet measurable ways.

(3) Uncovering Persuasive Strategies of Influencers and Media: Influence is not solely about reach but also about how messages are communicated. While prior studies often focus on structural or engagement metrics, few examine the rhetorical and persuasive mechanisms that drive audience alignment. This chapter applies classical rhetorical theory—ethos, logos, and pathos—combined with network contextualization, to analyze how political influencers and institutional news media craft persuasive messages. The study examines how rhetorical style and emotional framing influence audience composition to enhance engagement and shape collective opinions.

By systematically addressing these subproblems, this chapter advances a multi-dimensional understanding of online influence, integrating structural, temporal, emotional, and rhetorical perspectives. This approach

provides both methodological rigor and actionable insights for monitoring, understanding, and mitigating the mechanisms of digital persuasion, polarization, and opinion manipulation.

5.1 Identifying Key Influencers and Their Impact on Collective Opinion

This work focuses on the persuasive influence of individuals who possess the capacity to reshape the opinions and emotional climate of online communities. For this purpose, a robust framework was introduced for identifying influencers not merely by their popularity, but by their measurable ability to shift the opinions of their audience over time. Furthermore, an investigation was performed on the downstream effects of manipulating an influencer's opinion, analyzing how such interventions propagate through their audience and alter collective sentiment. Building on the FJ opinion dynamics model, the approach redefines influencers as those individuals whose presence measurably alters collective sentiment trajectories in their communities, especially when their expressed opinions are deliberately perturbed.
Once these influencers are identified, their strategic importance can be determined through a manipulation experiment. Specifically, the simulation shifts in opinion by selectively altering the initial sentiment values of key influencers and comparing the resulting impact on community-wide opinion evolution to that of perturbing randomly selected users. This allows us to assess whether targeted interventions — whether malicious (e.g., disinformation campaigns) or corrective (e.g., counter-radicalization messaging) — can produce disproportionate effects depending on who is influenced. This contribution is threefold:

- Influencer Identification: A method proposed to identify key influencers based on their dynamic, measurable impact on audience sentiment over time, going beyond traditional static metrics.
- Sentiment Manipulation Experiments: Opinion simulation conducted by manipulating influencer sentiment and assessing how

such changes propagate through the network compared to baseline (random user) manipulations.

- Propagation Analysis: The analysis of the structure of influence diffusion is performed, identifying both direct and indirect (second-degree) pathways through which influencer opinions change and reshape community sentiment.

5.1.1 Methodology

The core idea of the approach is to go beyond traditional centrality-based definitions of influence and instead define influencers as users whose initial opinions have the most significant impact on the final steady-state opinions of others in the network. By modeling the evolution of opinions through user interactions, it becomes possible to identify high-impact individuals and simulate how deliberate manipulation of their sentiments may influence broader ideological shifts within the community.
The methodology unfolds in several stages. First, a directed interaction graph is constructed based on user mentions, followed by community detection using modularity optimization. Next, opinion dynamics are simulated using the FJ model. The core of the influence identification approach is built on analyzing the fundamental matrix of the network's Laplacian structure. This allows us to measure how each user's initial opinion affects every other user's final opinion. An Influence Score is defined for each user and used to rank individuals, identifying the most influential actors within the network. To evaluate the actual impact of these users on opinion dynamics, a series of opinion manipulation experiments is conducted. Specifically, the study simulates how community-wide opinion shifts vary when the initial sentiments of top influencers are altered, compared to those of randomly selected users. By applying both positive and negative sentiment manipulations, the analysis assesses whether influential users possess a disproportionate capacity to steer overall community sentiment—either toward radicalization or moderation. Each component of this process is described in detail in the following subsections.

Graph Construction and Community Detection

This phase is dedicated to building an interaction graph $G = (V, E)$. In this case, E stands for the edges between users, and $V = \{v_1, v_2, \cdots, v_n\}$ indicates the users. The edges in the network are weighted by the frequency of mention between two nodes and are represented as $a_{(v_i, v_j)}$. Once the graph was constructed, the Louvain method [18] was employed to detect communities.

Opinion Dynamics Modeling

The FJ model was employed here to simulate the users' opinions within the network. This model assumes that each user embedded two opinions, the $x_i(t)$ which represent the expressed opinion at any t, and s_i represents the initial opinion on a particular subject. Crucially, s_i doesn't change during the procedure. The expressed viewpoint $x_i(t)$, on the other hand, changes over time as a result of interactions with different social network users. The following equation governs the evolution of these expressed thoughts at $(t + 1)$:

$$x_i(t+1) = \frac{s_i + \sum_{j \in N(i)} a_{ij} x_j(t)}{1 + \sum_{j \in N(i)} a_{ij}} \tag{5.1}$$

Where $N(i)$ represents the set of neighbors of the node.
Here, the graph created in this study was subjected to the FJ model, in which the average sentiment score (see the Data Preparation paragraph) was used as s_i. Each user's $x_i(t)$ is iteratively updated by the model until the opinions converge. The equilibrium vector indicates that this process has reached a stable state.

$$\mathbf{x} = (I + L)^{-1}.s \tag{5.2}$$

where I is the identity matrix, and L is the Laplacian matrix.
The matrix $(I + L)^{-1}$ highlights the structural relation within the graph, in which it demonstrates the impact of each s_i on equilibrium opinions.

Influence Identification Influencers are users whose opinions have a significant impact on those of others. This phase aims to find and rank the users by calculating their $InScore_i$ using (5.3):

$$InScore_i = \frac{1}{n} \sum_{j=1}^{n} \omega_{ij} \tag{5.3}$$

where ω_{ij} are the fundamental matrix. Users were rated according to this score to determine the most critical community influencers.

Opinion Manipulation In this step, two sentiment manipulations were conducted to investigate how changes in the influencer's opinion affect their audience in comparison to other regular users. The two simulations are as follows:

- Positive: $s_i = +1$ in this simulation, which means a significantly positive emotion.
- Negative: $s_i = -1$, indicating a very negative feeling.

The FJ model was run to compute a fresh set of equilibrium opinions, x_i, for each simulation. The $Avg(z)$ that represents the average equilibrium opinion was computed to examine the shifts in the opinions of influencers and random users. This measure captures the overall change in the sentiment of the network.

5.1.2 Experimentation

The experiments were conducted in two main phases. In the first phase, influential users within the detected communities were identified and ranked based on their influence scores derived from the opinion dynamics model. This phase enabled the isolation of individuals whose initial beliefs had the most significant impact on the equilibrium opinions of other network members. In the second phase, through deliberate opinion manipulation, the researchers examined how these influencers collectively shaped the community's sentiment. By altering the initial opinions of selected users—both influential and randomly chosen—the impact on

the final opinion landscape was simulated. The goal of this experiment was to determine whether manipulating the sentiment of top influencers resulted in significantly greater shifts in collective opinion compared to a baseline (random) manipulation. Results from both positive and negative sentiment perturbations were analyzed to assess the directional sensitivity and robustness of community opinion to such interventions.

Dataset

A publicly accessible dataset from Kaggle [64] was utilized in this work, comprising approximately two million tweets collected over one month. Rich metadata, including user identifiers, timestamps, tweet content, and engagement metrics such as the number of followers, likes, retweets, and replies, accompany each tweet in the dataset. This enables the construction of meaningful user interaction networks and facilitates sentiment-based opinion modeling. The preparation for the data includes the following:

- Data Cleaning: Removal of non-English and duplicate tweets.

- Interaction Extraction: Involves portraying user interaction by extracting the mention engagement from a tweet. Every time one user referenced another, a directed edge was formed between them, and the frequency of these mentions was calculated. To capture the degree and direction of user influence, these frequencies were used as edge weights in the interaction graph.

- Sentiment Analysis: A pre-trained sentiment analysis model based on RoBERTa was utilized to analyze users' tweets and assess their emotional sentiments. A sentiment polarity score was assigned to each tweet, and the average score was determined. In the opinion dynamics model, the user's initial opinion s_i was represented by this total score.

Rank	User	Batch	Influence Score	Retweets (Sum)	Number of Tweets	Followers Count
1	RealJamesWoods	Batch 1	0.8908	109251	8	2,685,154
2	w_terrence	Batch 1	0.8805	42741	15	1,188,925
3	Varneyco	Batch 1	0.8523	4800	55	663,854
4	Rasmussen_Poll	Batch 1	0.7605	2505	52	358,137
5	JudicialWatch	Batch 1	0.7009	42773	87	1,843,739
6	EpochTimes	Batch 2	0.5169	3232	36	319,635
7	trish_regan	Batch 2	0.4883	3152	6	737,555
8	WayneDupreeShow	Batch 2	0.3430	1385	20	504,846
9	Styx666Official	Batch 2	0.2417	417	6	90,474
10	realTrumpForce	Batch 2	0.1662	12682	88	87,096
11	RealMattCouch	Batch 3	0.1549	730	10	439,534
12	JoeTalkShow	Batch 3	0.1159	12369	110	107,473
13	Wizard_Predicts	Batch 3	0.1155	1967	57	12,659
14	Out5p0ken	Batch 3	0.1088	12	6	27,476
15	MarkSimoneNY	Batch 3	0.1001	13579	97	193,717

Figure 20: Detected Influencers

Influencer Identification

The method that was detailed in the methodology section was implemented in this experiment to detect and rank users based on their structural impact on opinion dynamics. All users were ranked in descending order of influence using the computed Influence Score ($InScore$).

- Batch 1: first 5 users.
- Batch 2: users ranked from 6 to 10.
- Batch 3: users ranked from 11 to 15.

Influential people are divided into several groups to examine the correlation between the Influence Score ($InScore$) and its actual impact on opinion dynamics within the network. This stratification allows us to observe whether users with higher $InScore$ induce more pronounced shifts in community sentiment compared to those with relatively lower scores. As summarized in Figure 20, the proposed FJ-based model demonstrated a notable advantage by not only detecting these users but also providing a fine-grained ranking through the computed $InScore$ values. This demonstrates FJ's ability to go beyond structural centrality and capture the nuanced impact of each user on the overall evolution of opinions in the network.

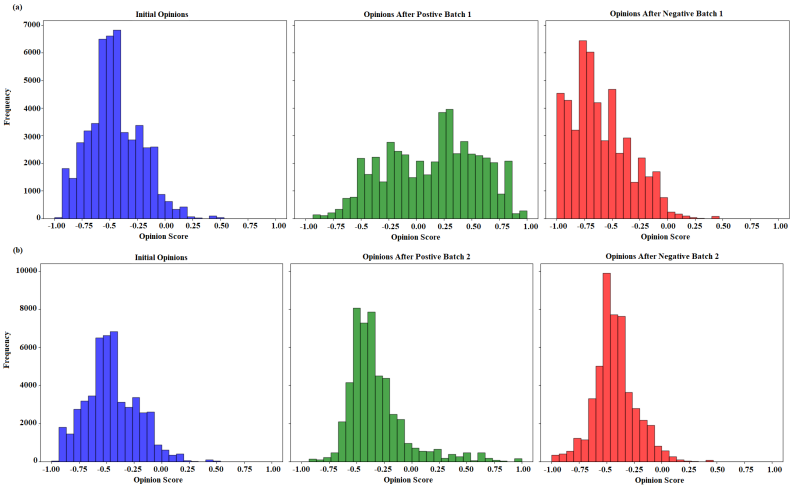

Figure 21: Batch 1 and Batch 2 opinions before and after manipulation.

Additionally, the relationship between the computed $InScore$ and key user engagement metrics was investigated. As shown in Figure 20, it reveals a high correlation between the$InScore$ and both interaction metrics, suggesting that users who receive more retweets and have larger audiences tend to exert greater influence on the opinion dynamics within the network. In contrast, the number of tweets produced by a user does not exhibit a meaningful correlation with $InScore$, indicating that mere content volume does not equate to influence. These findings underscore an essential distinction: high posting activity alone does not make a user influential. Instead, influence is more closely associated with the level of audience engagement and visibility, as reflected by the number of retweets and the size of the follower base. This analysis reinforces the validity of the $InScore$ metric by aligning it with meaningful engagement signals, while also highlighting its ability to distinguish between superficial activity and genuine opinion-shaping influence within social media ecosystems.

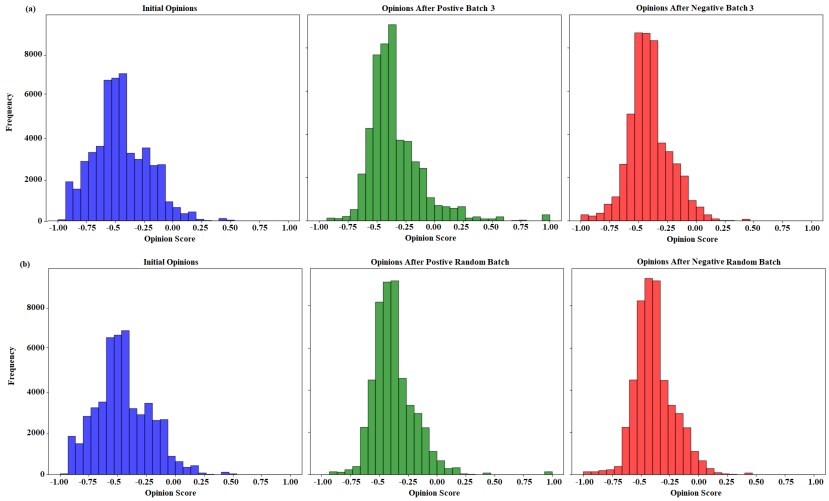

Figure 22: Batch 3 and random Batch opinions before and after manipulation.

Sentiment Manipulation

This experiment investigates how sentiment manipulation, specifically altering the initial opinions of selected users, impacts the overall opinion dynamics within the network. The goal is to assess whether influencers, as identified by their high $InScore$ values, exert a greater influence on the community's equilibrium opinion compared to users with lower rankings or those selected at random. To this end, both positive and negative sentiment manipulations were simulated on each user batch, and the resulting changes in the opinion at equilibrium were observed.

It can be observed that Batch 1 has a substantial impact on the network's collective sentiment. As displayed in Figure 21.a and Figure 23, the opinion change increased to 0.0866 in the case of positive manipulation, indicating a clear shift toward positivity. Conversely, negative manipulation further intensified polarization, dropping it to -0.5857 for influencers ranked 6–10, as shown in Figure 21. b illustrates more modest but still notable shifts. Positive manipulation resulted in an average opinion

Batch	Original Opinion	Positive Manipulation	Negative Manipulation
Batch 1	-0.3932	0.0866	-0.5857
Batch 2	-0.3932	-0.3408	-0.4291
Batch 3	-0.3932	-0.3647	-0.4120
Random Batch	-0.3932	-0.3904	-0.3940

Figure 23: Equilibrium Opinion Changes by Batch.

	Direct Connection		Indirect Connection	
	Batch 1	Batch 2	Batch 1	Batch 2
Positive Intervention	7	8	3	2
Negative Intervention	6	9	4	1

Figure 24: Sample of Users Affected after Manipulation.

change from -0.3932 to -0.3408, while negative manipulation decreased it to -0.429. On the other hand, Batch 3 demonstrated only a marginal influence. As shown in Figure 22.a and Figure 23, the average opinion experienced minimal variation under both positive and negative manipulations.

Finally, sentiment manipulation on random users yielded negligible changes. As seen in Figure 22. b (bottom row), the average network opinion shifted insignificantly from -0.3932 to -0.3904 under positive manipulation, and from -0.3932 to -0.3940 under negative manipulation—highlighting their limited influence on overall opinion dynamics. These results validate the hypothesis that users with higher $InScore$ values possess greater capacity to steer community sentiment, either toward optimism or deeper polarization. The correlation between $InScore$ and opinion-shifting power is evident: as $InScore$ increases, so does the magnitude of the impact on network-wide opinion.

Propagation Analysis

Following the identification of influencers and the evaluation of how their manipulated sentiments affect overall community opinion, this experiment investigates how influence propagates through the network. The propagation analysis aims to uncover the structural diffusion patterns triggered by opinion manipulation—specifically, to identify which users are most affected and whether the impact is confined to direct neighbors or also extends to indirect connections. To this end, the top users affected by both manipulations were examined for the first two influencer batches. This assessment focused on measuring how changes in the initial opinions of these influencers influence downstream nodes, distinguishing between direct neighbors (first-degree connections) and indirect neighbors (second-degree or higher connections).

As shown in Figure 24, Batch 1 influencers exhibited a clear and consistent pattern of wider influence propagation. Both direct and indirect neighbors were significantly affected under both manipulation types, indicating that their influence permeates the broader network. While the most impacted users are typically the influencers' immediate neighbors, a substantial number of indirectly connected users also experienced meaningful shifts in their equilibrium opinions. This demonstrates that top influencers have the capacity to disseminate influence beyond their immediate network, affecting multi-hop connections within the social graph. In contrast, the influence of Batch 2 users was found to be more localized. For both positive and negative manipulations, the majority of impacted users were direct neighbors. Specifically, 8 out of 10 users (positive manipulation) and 9 out of 10 users (negative manipulation) had direct ties to the manipulated influencer. This suggests a more constrained diffusion effect, likely due to lower $InScore$ values relative to Batch 1.

In summary, regardless of the manipulation type, this analysis reveals distinct propagation patterns tied to influencer rank. Manipulating the initial opinions of top-tier influencers (Batch 1) has a network-wide ripple effect, extending beyond immediate connections and in-

fluencing indirectly connected users. In contrast, lower-ranked influencers (Batch 2) primarily affect their immediate network. These findings underscore that as $InScore$ increases, the scope of influence broadens—from localized opinion shifts to widespread community-level sentiment changes—highlighting the systemic role of high-impact users in shaping digital opinion landscapes.

Hardware and Software Environment: All experiments were conducted using Google Colaboratory (Colab), a cloud-based computational platform provided by Google. The execution environment was equipped with an Intel Xeon CPU, approximately 12 GB of RAM, and an NVIDIA GPU (Tesla T4) when required for deep learning experiments. The operating system was a Linux-based environment managed by Colab.

5.2 Temporal and Emotional Dynamics of Influence in Online Communities

Understanding who truly shapes collective opinion in social networks requires moving beyond superficial measures of popularity or network position. This section addresses the critical problem of identifying and quantifying the actual influence of key users: which individuals can measurably shift community opinions, and how their structural position translates into genuine opinion change. For this purpose, a sentiment-aware framework based on the FJ opinion model [42] was employed to analyze how influence manifests, propagates, and varies across communities, thereby preparing the ground for subsequent investigations into temporal dynamics and emotional impact.

The approach is operationalized through reply-based social graphs constructed over multiple time frames, community detection via METIS, influencer selection, and finally, the analysis of opinion dynamics. This approach exploits two scenarios with and without the presence of influencers to causally infer their impact on both community opinion trajectories and emotional atmospheres. The reported experiments of this model to real-world Twitter data on vaccine discourse reveals key in-

sights: first, that topological prominence alone does not predict true opinion-shaping power; second, that the emotional signals embedded in influencer content (e.g., fear or anger) strongly align with and possibly drive community-level emotional shifts; and third, that influence is not monolithic even similarly ranked influencers can project distinct emotional patterns and thus play divergent roles in shaping collective sentiment.

5.2.1 Methodology

This section introduces the methodology for analyzing how influencers dynamically shape opinion and emotional trajectories within online communities. A multi-phase method is proposed that integrates social network analysis, graph-based influencer detection, and sentiment-aware opinion modeling grounded in the FJ framework. The core idea behind this methodology is to move beyond static representations of influence and instead model how influence manifests and evolves, particularly in relation to the emotional impact that influencers exert on their surrounding communities. The approach captures this dynamic by comparing opinion formation in two scenarios: one in which selected influencers are embedded within the network and another in which they are removed. By observing the divergence between these two configurations across successive time frames, the evolving impact of influencers on both community-level opinion states and the emotional climate of discourse is causally inferred.

The methodology consists of five sequential components, as shown in Figure 25: (1) Data Processing, where user-generated content is cleaned and analyzed for sentiment; (2) Social Network Construction, where user interactions are modeled as temporal reply graphs; (3) Community Detection, which partitions the network into distinct topical clusters using the METIS algorithm; (4) Influencer Selection, where key actors are identified using normalized hybrid centrality metrics across multiple time frames; and (5) Sentiment-aware Opinion Modeling, which simulate how initial sentiment (encoded as user opinion) propagates under

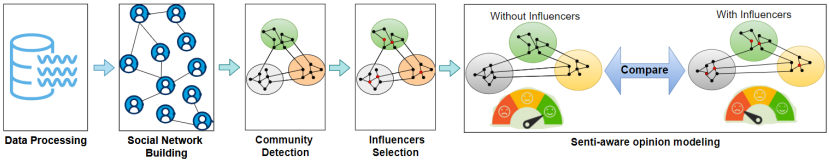

Figure 25: Data pipeline.

social influence, accounting for both temporal interactions and individual stubbornness.

Network Construction & Community Detection

The first step is dedicated to constructing a graph that encompasses nodes (users) and the edges (reply interactions) among them for each time frame t^{th}. Here, the edges are weighted by the reply frequency between the two users. Following the graph construction, the METIS algorithm was used to perform community detection at each t^{th}.

Influencers Selection

This step aims to identify the influencers within each community. For this purpose, degree centrality and PageRank centrality metrics were used. Both metrics were computed for each user and then averaged. Any user with a value greater than 0.5 was identified as an influencer. Finally, only the influencers present in all t^{th} time steps were taken into account.

Sentiment-aware Opinion Modelling

This step constitutes the core of the methodology, where the FJ model is employed to simulate the evolution of opinion in each community. This model is represent by Equation (5.4)

$$\mathbf{x}(t+1) = \mathbf{W}^t\mathbf{x}(t) + (\mathbf{I} - \mathbf{W}^t)\mathbf{s} \tag{5.4}$$

Here, $\mathbf{x}(t)$ is the opinon at t, $\mathbf{s}$ is the stubbornness levels, $\mathbf{I}$ is the identity matrix, and $\mathbf{W}^t$ is the influence matrix at t.
Equation (5.4) is applied at each t^{th}. For each t^{th}, it is repeated, where opinions computed at time t become the starting opinions for time $t+1$. This approach allows observing how initial sentiments evolve based on modeled social interactions.
In this approach, the user's opinion is assumed to represent the emotional reaction; the RoBERTa model [75] was employed to extract sentiment scores for each tweet. These sentiment scores are aggregated per user to represent their emotional perspectives over time, hereafter referred to as senti-aware opinions. Thus, the initial opinions are encoded in $\mathbf{x}(0)$. W^t is an matrix quantifies the influence weight of j on i at t. The rows are normalized such that $\sum_j w_{ij}^t = 1$. Furthermore, a uniform stubbornness score is assumed, where historical sentiment scores are encoded as $s_i = 1$ for all users, indicating complete stubbornness and no alteration of initial opinions. Finally, to investigate the impact of influencers, this approach was applied to two networks: one with influencers and the other without them. By comparing the evolution of opinions in these two scenarios across the four time frames, the impact of influencers on opinion dynamics was sequentially assessed.

5.2.2 Experimentation

Dataset & Data preparation

To empirically validate the proposed framework, the Twitter Vaccination Dataset[1] was employed, a publicly available collection of tweets related to COVID-19 vaccination discourse. The dataset includes all tweets containing the keyword "vaccination," offering a rich source of public opinion and user interaction data. The dataset contains more than 300,000 tweets posted by 50,000 unique users. Each entry includes not only the tweet content but also valuable metadata such as the timestamp of publication, user identifiers, follower and following counts, and geolocation

[1] https://www.kaggle.com/datasets/keplaxo/twitter-vaccination-dataset/data

data (when available). This comprehensive structure allows for both temporal and structural modeling of user interactions. To facilitate longitudinal analysis of opinion dynamics, the dataset was partitioned into four equal temporal subsets based on tweet volume: Q1, Q2, Q3, and Q4, each corresponding to three months. This temporal segmentation supports the construction of time-evolving interaction graphs, facilitating the dynamic tracking of influencer roles and sentiment propagation across communities.

Before conducting the main experiments, the dataset underwent a series of preprocessing steps as outlined in the Methodology section. These steps included graph construction, community detection, and influencer identification, all of which form the foundation for subsequent opinion and emotion modeling. User interactions were modeled as reply-based graphs. This process produced four distinct weighted graphs, each corresponding to one of the temporal segments (Q1–Q4) defined in the dataset. These graphs capture the evolving topology of user interactions over time. Following the construction of the graph, the METIS algorithm was applied to detect communities within each temporal graph. METIS partitions the network into coherent subgroups by optimizing for intra-community density, enabling the isolation of clusters of users who engage with similar topics or narratives. For each time frame, two major communities were identified, allowing comparative analysis across both time and group structure. Once the communities were established, influential users within each community were identified using a hybrid centrality-based method detailed in the methodology, which combined degree centrality and PageRank. These influencers serve as the focal points for the dynamic analysis of opinion and emotional influence.

Preliminary Sentiment Analysis

First, the distribution of opinion within each community is investigated to contextualize the dynamics observed in later stages. From Figure 26, it can be seen that the sentiment distribution is balanced for community 0, indicating that it comprises heterogeneous perspectives and varying stances on the vaccination topic, which reflects a pluralistic discourse

space. In contrast, Community 1 demonstrates a marked skew toward negative sentiment. The sentiment distribution reveals a dominant concentration of pessimistic or critical views, indicating a more ideologically aligned group with strong opposition to vaccination narratives. This homogeneity hints at the presence of an echo chamber, where reinforcing sentiments may intensify polarization and reduce exposure to countervailing viewpoints.

Further analysis of influencer characteristics, as detailed in Figure 27, underscores a striking asymmetry in influence concentration across the two communities. Community 1 not only hosts a greater number of influencers but also includes users with substantially higher influence scores and engagement metrics (e.g., retweets, likes). The correlation between influence scores and interaction volume suggests that these users play a central role in shaping discourse, mobilizing opinions, and amplifying emotional tone within their community. Taken together, this preliminary analysis highlights two distinct community configurations: Community 0, characterized by a diverse and diffuse opinion structure, and Community 1, marked by concentrated influence and a prevailing negative sentiment. These structural and emotional differences provide critical context for the subsequent modeling of opinion evolution and emotional influence, and underscore the necessity of temporally sensitive, sentiment-aware frameworks for capturing the dynamics of digital influence.

Hardware and Software Environment: All experiments were conducted using Google Colaboratory (Colab), a cloud-based computational platform provided by Google. The execution environment was equipped with an Intel Xeon CPU, approximately 12 GB of RAM, and an NVIDIA GPU (Tesla T4) when required for deep learning experiments. The operating system was a Linux-based environment managed by Colab.

Results on Opinion

To assess the temporal influence of key actors on community sentiment, the Senti-aware opinion modeling framework was applied under two

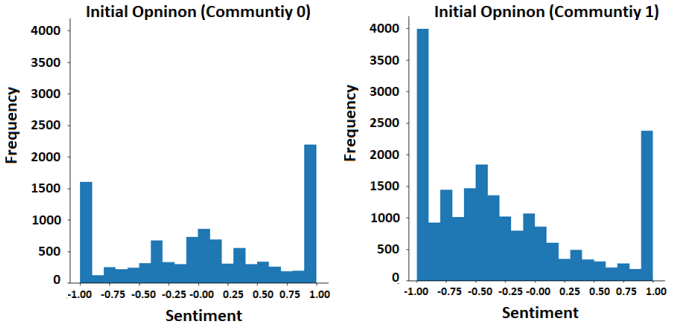

Figure 26: Sentiment analysis for the detected communities.

Twitter User	Number of Tweets	Retweets Count	Likes Count	Influence Score	Community
dortimi	45,369	33,937	502,687	0.88	1
mcfunny	39,569	30,157	490,624	0.79	1
kidoctr	670	1,056	39,562	0.64	1
gavi	5,074	1,569	90,698	0.62	0
nytimes	2,062	2,032	75,980	0.61	1
thereal_truther	376	106	26,044	0.58	1
ianfmusgrave	1,056	55	46,521	0.55	1
badzoot7	450	61	36,812	0.52	1

Figure 27: Detected influential users.

comparative conditions: one in which identified influencers were retained in the interaction graph, and another in which those influencers were removed.

Figure 28 presents the results for Community 0, focusing on the sole influencer ”gavi.” In the presence of ”gavi” (Figure 28a), the influencer maintains a relatively stable and higher opinion trajectory than the community average, which exhibits greater variability. However, in the absence of ”gavi” (Figure 28b), the community’s opinion curve remains essentially unchanged, with only minor deviations observed in quarters Q3 and Q4. This outcome suggests that although ”Gavi” maintained a distinctive opinion, its presence did not exert a significant directional influence on the broader community’s sentiment over time.

By contrast, Figure 29 illustrates the opinion evolution in Community 1, which contained a higher density of influential users. In Figure 29(a), a relatively stable community opinion is observed, while the seven iden-

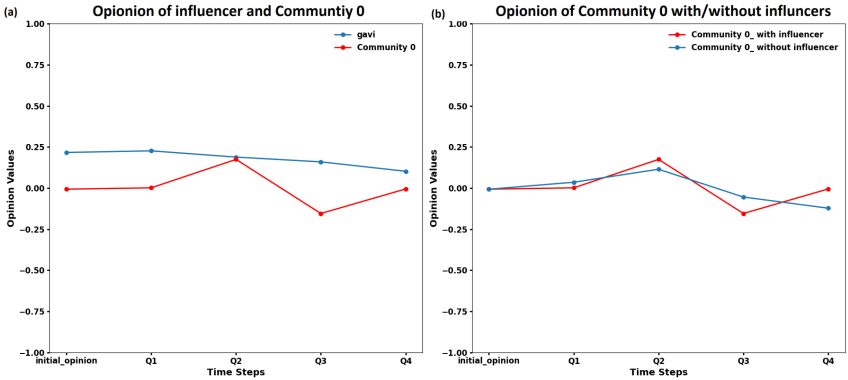

Figure 28: The opinions of community0 with/out influencers.

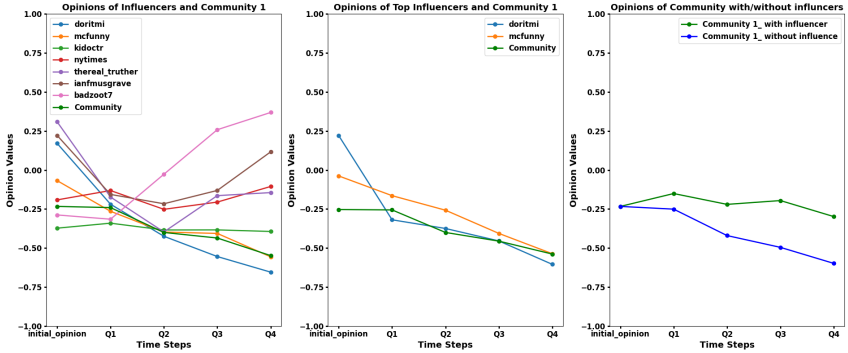

Figure 29: The opinions of community1 with/out influencers.

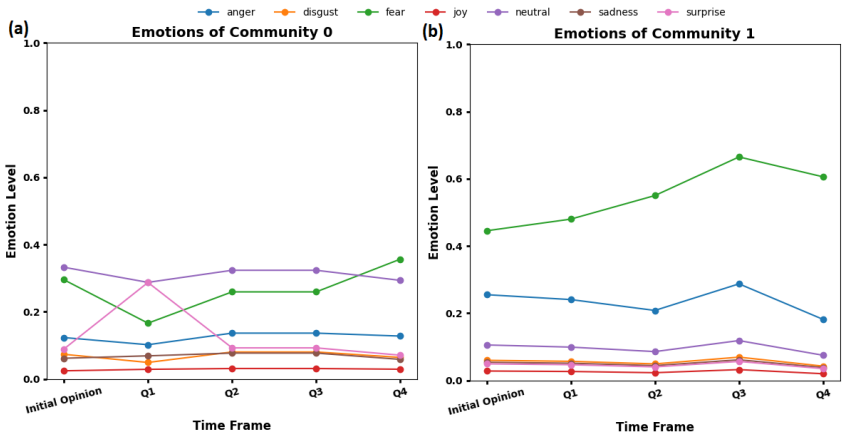

Figure 30: Subfigures (a,b) represent the variation of community 0 and community 1 emotion.

tified influencers demonstrate varied and individualized trajectories. However, Figure 29(b) isolates the two top influencers—"dortimi" and "mcfunny," both with influence scores above 0.8 and reveals their outsized impact relative to the other five influencers, whose presence yields minimal effect. These top influencers consistently steer the community's opinion toward more negative sentiment, reinforcing the idea that influence is not evenly distributed among structurally central users. Finally, Figure 29(c) shows the outcome of removing all influencers from the Community 1 graph and applying the FJ. It can be seen that the opinion curve is notably more stable and consistently lower than in the previous conditions. The absence of high-impact influencers flattens the emotional volatility of the group, confirming that "dortimi" and "mcfunny" played a substantial role in driving both the intensity and direction of opinion evolution. The influence is a dynamic and asymmetric phenomenon, where only a select subset of actors exerts a sustained and measurable impact on community sentiment. This insight cannot be captured solely through static network metrics.

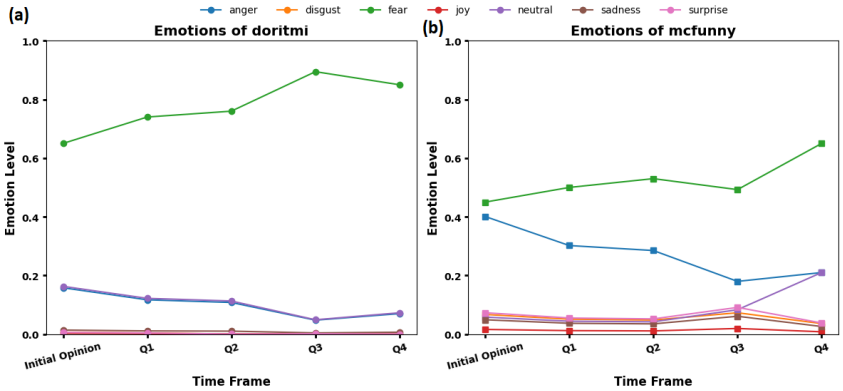

Figure 31: Comparison of Dortimi and McFunny emotion.

Results on Emotion

This section focuses on how emotional expressions propagate and evolve.

Figure 30 presents the temporal evolution of emotional states in Community 0 and Community 1 across five stages. In Figure 30(a), emotional expressions show considerable fluctuation over time for Community 0, with no single emotion dominating, indicative of a heterogeneous, emotionally diverse community. In contrast, the Figure 30(b) for Community 1 displays a markedly different emotional profile. Two emotions stand out as most prominent. Fear rises notably from 0.4 at the start to above 0.6 by Q3, then dips slightly in Q4, whereas Anger shows a consistent decline from 0.35 to under 0.25. This emotional polarization signals a shift in how discourse is emotionally framed within the community. Given this pattern, a further investigation was conducted into the emotional behavior of the two top influencers in Community 1: "dortimi" and "mcfunny," who were previously shown to have an impactful. Their emotion trajectories are depicted in Figure 31. For "dortimi" (Figure 31.a), Fear shows a steep upward trend in Q4, while Anger remains low and declines further. In contrast, "mcfunny" (Figure 31.b) exhibits a more mixed emotional profile, with Fear increasing from

0.35 to 0.6 and Anger showing moderate expression before tapering off toward Q4.

This analysis reveals two important insights. First, the alignment between the emotional states of top influencers and the emotional evolution of the communities they inhabit, particularly in the case of Fear. Second, and more crucially, even influencers with comparable influence scores (as shown in Figure 27) do not project identical emotional patterns. "Dortimi" presents a clear and consistent emotional narrative, dominated by Fear, while "McFunny" exhibits a more diffuse emotional profile, blending both Fear and Anger. These findings suggest that influence is not solely a function of network centrality or volume of interaction, but also of emotional framing—how specific emotions are communicated and absorbed within the social fabric. The same quantitative influence score can mask qualitatively different emotional effects, potentially leading to divergent community responses and downstream behavioral consequences.

5.3 Rhetorical and Persuasive Mechanisms of Influencers and News Media

Beyond who influences others and how opinions evolve, influence also depends on how messages are communicated. This section addresses the problem of identifying and comparing the persuasive strategies of political influencers and institutional news media, focusing on the rhetorical and emotional tactics embedded in their language. Social media has become a rhetorical battleground where political influencers, unlike institutional news media, use the rhetorical language as a tool to engage in agile and audience-aligned discourse by embedding it within their posts and multimedia content [24]. This work addresses two central research questions: (1) How do political influencers and news media differ in their use of persuasive strategies, and how does this impact audience engagement? (2) How do influencers and news media accounts adapt their persuasive strategies across their audiences? These questions drive

the investigation into the stylistic and contextual elements that govern digital persuasion. To address these issues, the analytical lens is shifted toward the linguistic and stylistic dimensions of user-generated content. Anchored in Aristotle's classical rhetorical appeals—ethos (credibility), logos (logic), and pathos (emotion)—the analysis examines how persuasive strategies vary between political influencers and institutional news accounts, and how these differences manifest in rhetorical alignment, emotional tone, and discursive resonance with audiences. This approach integrates rhetorical theory, sentiment modeling, and ego-network contextualization to move beyond static content analysis and uncover how persuasive styles dynamically adapt to audience composition and network structure. The methodology begins with the construction of a directed mention graph, from which high-engagement political influencers and news media accounts are identified. Ego networks are then extracted for a curated sample of users in both groups to study how persuasion functions within their immediate social spheres. This design enables the capture of how persuasive strategies vary across user types and evolve in response to local network context.

Through a large-scale analysis of Twitter interactions, it was found that while institutional news media retain formal authority, they often suffer from rhetorical distance, exhibiting low linguistic alignment with audience discourse. In contrast, political influencers demonstrate greater rhetorical flexibility, emotional attunement, and stylistic mimicry, enabling them to amplify engagement and reinforce polarization more effectively. By capturing these rhetorical differentials, this contribution advances the chapter's overarching goal: not only to understand who influences and how influence spreads, but to uncover how language itself becomes a vehicle of persuasion, polarization, and digital control.

5.3.1 Methodology

The methodology is designed as a multi-stage process that combines rhetorical modeling, sentiment analysis, and network-based contextualization. The core objective of this approach is to move beyond traditional

structural analysis and capture the stylistic, emotional, and adaptive dimensions of persuasion as they unfold in real-time social media interactions. The methodology proceeds in four stages: (1) Data Preprocessing, (2) User Identification, (3) Persuasion Strategy Detection, and (4) Ego-Network Contextualization.
Before detailing the methodological pipeline, an overview of the foundational concepts of Aristotle's classical appeals is presented. This theoretical grounding provides the interpretive framework through which persuasive strategies are identified, categorized, and analyzed.

Foundations of Rhetoric Theory

Rhetoric theory is defined as the art of persuasion, originating in the practices of public discourse in ancient Greece. It focuses on how language can be used to convince audiences [92]. One of the earliest frameworks that studied rhetorical theory is Aristotle's framework. Aristotle considers that the effective persuasive communication often integrates three primary modes of appeal: Ethos (credibility), Pathos (emotion), and Logos (logic and reasoning) [106]. Table 17 shows examples of social media posts categorized into logos, ethos, and pathos.

- Ethos: This mode refers to the appeal to trustworthiness, authority, and credibility of the speaker. Aristotle assumes that the integrity and expertise of the speaker significantly impact the persuasiveness of their argument. On social media platforms, key influencers establish trust and legitimacy through various signals, including identity verification, affiliations (such as partnerships with news organizations or academic institutions), and other indicators of credibility.

- Logos: Logos mode refers to the appeal of a logical speaker, which is often grounded in rational argumentation. This mode relies on evidence and facts to persuade the audience. In digital contexts, logos can be presented through data visualizations, links to trustworthy sources, or well-structured argumentative threads.

Table 17: Examples of Persuasion Types (Ethos, Logos, Pathos) in social media posts.

Persuasive Type	Example Post	Explanation
Ethos	*"As a verified journalist at @NewsOrg, I assure you that these facts are accurate and have been thoroughly checked."*	Demonstrates credibility through identity verification and professional affiliation.
Logos	*"According to the latest CDC report, vaccination reduces COVID-19 risk by 90%. See full data here: [link]."*	Uses data and evidence to support the argument logically.
Pathos	*"Seeing children suffer from this preventable disease breaks my heart. We must act now to protect them! emojiheart, emojicry"*	Appeals to emotions through language and emojis to evoke sympathy and urgency.

- Pathos: Pathos mode involves the use of emotional language to bring out feelings such as anger, fear, or pride in the audience. Posts and messages with their multimodal affordances (e.g., images, videos, and emojis) can serve as a ground for Pathos appeals.

Understanding these modes together provides a robust theoretical framework for identifying how persuasive language influences opinion and behaviours.

Dataset

A public dataset from Kaggle [64] was used in these experiments, which was collected from the X (formerly Twitter) platform. The dataset collected using the #Trump and #Biden hashtags encompasses more than two million tweets. Each tweet contains various attributes, including the tweet timestamp, user location, number of retweets, likes, follower count, and other relevant details. For the analysis, the focus was placed

exclusively on posts from users located within the United States. The final dataset consists of 26,138 tweets related to the U.S. election, posted by thousands of unique users.

Content analysis

An analysis was performed on all users' tweets in the dataset to assess the emotional tone of each post through sentiment scoring, and second, to identify and label the specific persuasive strategies embedded within the posts based on Aristotle's theory. For this purpose, an LLM technique was employed to extract sentiment scores and identify the persuasion techniques used in each post.

Sentiment Analysis To capture the emotional tone embedded in user content, a sentiment analysis technique leveraging a pre-trained RoBERTa model [75] was applied. Specifically, the textual data were first preprocessed by cleaning and tokenizing each post to meet the model's input requirements. The processed text was then fed into the RoBERTa model for sentiment classification, which outputs a probability distribution across sentiment classes: negative, neutral, and positive. These sentiment scores serve as proxies for the emotional leaning of each post and were subsequently aggregated across multiple posts to assess the overall opinion orientation of individual users.

Persuasion strategies Annotation To identify the rhetorical modes within tweets, a fine-tuned multilingual transformer model [22] developed based on SemEval 2024 (Task 4) was employed. This model was trained on an annotated dataset containing various linguistic contexts and is capable of detecting a wide range of persuasive techniques. Given a post, the model outputs one or more persuasion strategy labels such as appeal to emotion, causal oversimplification, authority transfer, and loaded language. For the analysis, these labels were consolidated into Aristotle's classical rhetorical categories: ethos, pathos, and logos. This mapping was performed based on the communicative intent of each strategy. For instance, techniques such as appeal to emotion and

flag-waving were grouped under pathos due to their affective emphasis; strategies like appeal to authority and testimonial were categorized as ethos, reflecting their reliance on credibility and character; and logic-driven techniques such as causal oversimplification and straw man were assigned to logos, capturing their argumentative structure.

Graph Construction & Ego Network

In this stage, a graph $G = (V, E)$ was constructed to model user interactions based on textual content. In this graph, V represents a unique user, while the edge from one node to another, E, denotes a mention between these nodes, i.e., an explicit reference from one user to another within a post. Mentions were extracted by parsing the text of each user-generated post (e.g., using the "@username" format). To enrich the constructed graph, the edges between nodes were weighted. Specifically, for each pair of users, the cosine similarity between the content of V_i and V_j tweets was computed, and the average cosine similarity of all tweet pairs between these two users was assigned as the edge weight.

To determine the leading actors within the interaction graph, the focus was first placed on identifying influencer accounts by implementing a detection model used in [13]. Specifically, the FJ model was employed, which estimates user influence based on iterations of opinion propagation and stability. The top five users with the highest influence scores, as determined by the FJ model, were selected as key influencers. Simultaneously, the five most significant news media accounts were identified by calculating engagement metrics—such as total mentions, retweets, and responses received—which reflected their prominence within the social network.

After identifying the main actors within the interaction graph, the ego networks for each influencer and news media account were extracted. An ego network subgraph includes a central node (the "ego") and all nodes directly connected to it (the "alters"), along with the connections

between these alters. This localized network structure provides an examination of how persuasion strategies change in relation to the audience.

5.3.2 Experimentation

The persuasion strategies used by influencers and news media accounts(RQ1)

To address RQ1, a comparative analysis was conducted on the distribution of persuasive strategies—ethos, pathos, logos, and non-persuasive language—across posts made by political influencers and news media accounts. Generally, the study found that political influencers and news media employ different persuasive strategies. As shown in Table 18 and Figure 32, the results show that the influencer accounts discussing the election on the X platform often emphasize credibility and employ language characterized by logical appeals and fact-based argumentation to engage their audiences effectively. In contrast to the previous work by [130], which found that political leaders strongly rely on emotional appeals (Pathos) to connect with audiences, the analysis reveals a different rhetorical orientation among influencers. Specifically, 41.2% of their posts utilize ethos-based strategies, while 47.8% rely on logos-driven content. Conversely, news media accounts predominantly adopt ethos as their primary rhetorical strategy, with 91.5% of their posts falling into this category. This suggests that credibility and institutional authority appear to be the default rhetorical mode for news media accounts, likely reflecting the normative standards of journalistic objectivity.
Notably, emotional appeals (pathos) are absent in influencer content and appear only marginally at 1.1% in news media posts, which suggests a general reluctance across both groups to employ affective rhetoric. Beyond the three classical modes of persuasion, a subset of posts from both influencers and news media accounts contain no persuasive language at all. Specifically, 11.0% of influencer posts and 6.9% of news media posts fall into this non-persuasive category, highlighting the presence of con-

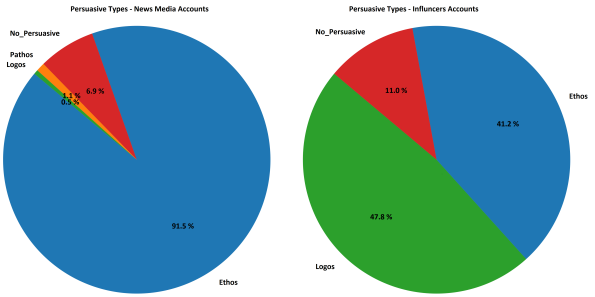

Figure 32: Proportional Use of Ethos, Logos, Pathos, and Non-Persuasive Content in Influencer vs. News Media Posts.

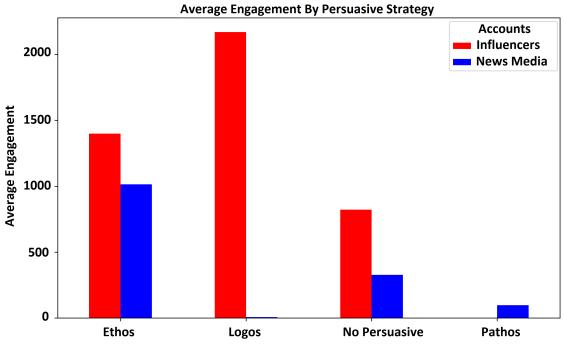

Figure 33: Comparison of Total Engagement Between Influencers and News Media Accounts.

Table 18: Distribution of Persuasion Types in Influencers and News Media Posts.

Posts Category	Influencers	News Media
Ethos posts	87	173
Logos posts	75	1
Pathos posts	0	2
No Persuasive posts	20	13
Total number of posts	182	189

tent that is either purely informational or strategically neutral in tone.
Building upon this analysis of rhetorical distribution, Figure 33 illustrates how these persuasive strategies correspond to audience engagement across both political influencers and news media accounts. The results reveal a marked disparity in impact, with influencers consistently generating higher average engagement per post, particularly when employing ethos- and logos-based appeals. In contrast, news media posts, despite their heavy reliance on ethos, typically yield more subdued levels of audience interaction. This suggests that while both groups employ credibility-based appeals, influencer content tends to resonate more actively with audiences.

How do influencers and media accounts adapt their persuasive strategies across different audience segments(RQ2)

To address RQ2, an ego network analysis was conducted on both political influencers and news media accounts to study how their persuasive language aligns with the prevailing discourse within their immediate network environment. This approach aimed to assess whether key actors exhibit linguistic homophily, that is, a tendency to communicate in ways that mirror the rhetorical style of their direct connections.
To quantify this alignment, a homophily index was employed to measure the degree of textual similarity between each central (ego) user and their corresponding network of interlocutors (alters). The homophily index, defined in Equation 5.5, is computed as the proportion of an ego's neighbors whose language demonstrates high semantic similarity (co-

sine similarity ¿ 0.7) with that of the ego user:

$$\text{Homophily Index} = \frac{M}{N} \tag{5.5}$$

Where:

- N: The total number of users within the ego's immediate network (neighbors)
- M: The number of those neighbors whose language exhibits a cosine similarity greater than 0.7 with that of the ego

Table 19: Homophily index for influencer and news media ego networks

Accounts	Persuasive Type	Homophily Index
Influencers	Ethos	71.2%
	Logos	76.5%
News Media	Ethos	22.7%

As demonstrated in Table 19, the findings reveal that political influencers exhibit high levels of rhetorical homophily within their ego networks, regardless of the specific persuasive strategy employed; this outcome stands in contrast to the patterns observed among news media accounts. Influencers who predominantly use ethos-based rhetoric show a homophily index of 71.2%, while those who employ logos-based appeals achieve an even higher index of 76.5%. In contrast, news media accounts using ethos demonstrate a significantly lower homophily index of only 22.7%, indicating minimal linguistic alignment between their rhetorical style and that of their immediate audience. These results suggest that influencers, whether appealing through credibility (ethos) or logic (logos), demonstrate a high degree of rhetorical synchronization with their audiences, reflecting a strong influence between influencer and audience. Conversely, news media accounts, despite their heavy reliance on ethos, appear discursively detached from their immediate networks, underscoring a potential disconnect between institutional communication styles and the language of their audiences.

5.4 Discussion

This chapter presents a multi-faceted, empirically grounded exploration of human-driven influence and digital persuasion in online social networks. Particularly, deliberate actions of influential actors and their language contribute to the radicalization and polarization pathway within the social network.

The results of this chapter provide robust evidence that influential actors significantly shape opinion dynamics in online communities, not only due to their network positions but also through their effective rhetorical strategies. By integrating the Friedkin–Johnsen opinion dynamics model with sentiment and rhetorical analysis, the study demonstrates that a small subset of influencers can induce substantial opinion shifts across communities, with their impact extending beyond direct followers to secondary connections. Moreover, a comparative analysis of influencers and news media reveals distinct patterns of persuasion: influencers often rely on hybrid strategies that blend credibility (ethos) and logical appeals (logos). In contrast, news media predominantly employ ethos-driven communication, which tends to have lower resonance among audiences.

These findings contribute to ongoing debates in communication and network science regarding the mechanisms of influence in digital environments. Whereas prior work has often measured influence through visibility or centrality metrics, this chapter demonstrates that influence must also be understood through the lens of rhetorical adaptation and emotional resonance. The strong effects of influencer interventions observed in simulation experiments lend support to theories of opinion cascades and echo the role of "persuasive elites" in shaping discourse within fragmented networks. By showing how influencers can function as accelerants of polarization or as potential moderators depending on their strategies, this study adds nuance to the literature on digital persuasion and political communication.

The practical implications are equally significant. For platform governance, the results highlight the importance of monitoring not only the

structural prominence of actors but also their discursive strategies. Policymakers and civil society organizations can leverage these insights to design interventions that amplify the influence of constructive voices and mitigate the reach of those promoting extremist narratives. Furthermore, the evidence that influencers exert influence beyond their immediate networks suggests that interventions targeting a small set of actors may have disproportionate effects on overall discourse.

In conclusion, this chapter directly responds to RQ3 by demonstrating how influential actors shape opinion dynamics through both network positions and rhetorical strategies, thereby realizing the thesis's third contribution on the human dimension of online radicalization.

Chapter 6

Conclusion and Future Work

This thesis examined online radicalization in social networks as an emergent phenomenon arising from the interaction of user behavior, algorithmic curation, and persuasive actors, and frames this process as a cognitive security and information integrity threat that undermines societal resilience and trust in digital information ecosystems. To address this, it employed a multi-method approach that integrates behavioral analysis, network science, and machine learning. Three core findings emerge:

First (RQ1), this thesis shows that echo chambers can be detected and monitored more effectively when behavioral and affective indicators are combined with network embeddings. These indicators not only identify ideologically insulated communities but also significantly predict whether individual users are likely to remain in or depart from such communities. Emotional polarization and toxic language are consistently correlated with higher cohesion within echo chambers.

Second (RQ2), the investigation advanced to a higher level, demonstrating that recommender systems contribute to ideological enclosure

through reductions in content diversity; however, structural interventions can mitigate this process. The diagnostic diversity indicator proposed here captures latent tendencies toward radicalization in recommendation flows, and the DRLGR reinforcement-learning rewiring approach reduces radicalization metrics while maintaining engagement, demonstrating that mitigation need not come at the expense of user relevance.

Third (RQ3), the focus shifted to studying the role of influential actors in shaping audience opinion. This thesis demonstrates that a small set of influential actors can produce disproportionate shifts in opinion. Influence depends on both structural position and discursive alignment: influencers who adapt rhetorical style to audience registers generate greater resonance and broader downstream effects than institutional accounts that rely primarily on ethos.

Building on these findings, several directions of exploration are needed for future work: First, further development of methodologies for real-time detection and mitigation of radicalization is essential, as timely identification of harmful trajectories would enable proactive interventions before communities become deeply entrenched. Second, cross-platform and multimodal validation should be pursued to ensure the generalizability and robustness of the proposed indicators and mitigation strategies, extending analyses beyond text to include video, audio, and image modalities that increasingly shape online discourse. Third, translating the theoretical frameworks and experimental findings into actionable applications through deployable systems, APIs, or pilot programs with industry partners will be critical for bridging the gap between academic research and practice. Such deployments would allow for live testing of mitigation strategies, the evaluation of user experience and engagement, and the refinement of tools in collaboration with platforms and policymakers.

Consequently, these objectives, together, advance the state of the art by providing frameworks that offer practical tools for measuring and miti-

gating radicalization. Although the contributions provided by this thesis, several directions of exploration are needed for future work. First, further exploration into methodologies that can be employed for real-time radicalization detection and mitigation is essential. These real-time investigations could enable proactive identification of harmful trajectories and timely deployment of countermeasures. Another direction should focus on translating the theoretical frameworks into actionable applications. This involves developing deployable systems or APIs that integrate the findings into platform operations. These types of applications can support a range of stakeholders, from tech companies aiming to implement safer design practices.

Appendix A

Appendix Title

Bibliography

[1] Zakia Acharoui et al. “Identifying Political Influencers on YouTube During the 2016 Moroccan General Election”. In: *Procedia Computer Science* 170 (2020), pp. 1102–1109.

[2] Swati Agarwal and Ashish Sureka. “Topic-Specific YouTube Crawling to Detect Online Radicalization”. In: *Databases in Networked Information Systems: 10th International Workshop, DNIS 2015, Aizu-Wakamatsu, Japan, March 23-25, 2015. Proceedings 10*. Springer. 2015, pp. 133–151.

[3] Faisal Alatawi, Paras Sheth, and Huan Liu. “Quantifying the Echo Chamber Effect: An Embedding Distance-Based Approach”. In: *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. 2023, pp. 38–45.

[4] Saja Aldera et al. “Exploratory Data Analysis and Classification of a New Arabic Online Extremism Dataset”. In: *IEEE Access* 9 (2021), pp. 161613–161626.

[5] Henrique Ferraz de Arruda et al. “Modelling How Social Network Algorithms Can Influence Opinion Polarization”. In: *Information Sciences* 588 (2022), pp. 265–278.

[6] Henrique Ferraz de Arruda et al. “Modelling How Social Network Algorithms Can Influence Opinion Polarization”. In: *Information Sciences* 588 (2022), pp. 265–278.

[7] Peddi Gowtham Balaji et al. “Cyberbullying Detection on Multiclass Data Using Machine Learning and a Hybrid CNN-BiLSTM Architecture”. In: *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*. Vol. 1. IEEE. 2024, pp. 1–6.

[8] Vandna Batra and Suresh Kumar. "A Semi-Automated Hybrid Approach to Identify Radicalization on Social Digital Platform". In: *Indonesian Journal of Electrical Engineering and Computer Science* 27 (July 2022), p. 563. DOI: 10.11591/ijeecs.v27.i1.pp563-572.

[9] L Beckett and JC Wong. "The Misinformation Media Machine Amplifying Trump's Election Lies". In: *The Guardian* 10 (2020).

[10] Alessandro Bellina et al. "Effect of Collaborative-Filtering-Based Recommendation Algorithms on Opinion Polarization". In: *Physical Review E* 108.5 (2023), p. 054304.

[11] Yochai Benkler, Robert Faris, and Hal Roberts. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press, 2018.

[12] Omran Berjawi, Giuseppe Fenza, and Vincenzo Loia. "A Comprehensive Survey of Detection and Prevention Approaches for Online Radicalization: Identifying Gaps and Future Directions". In: *IEEE Access* 11 (2023), pp. 120463–120491.

[13] Omran Berjawi et al. "Dynamic Analysis of Influencer Impact on Opinion Formation in Social Networks". In: *International Conference on Web Information Systems Engineering*. Springer. 2024, pp. 394–408.

[14] Omran Berjawi et al. "Understanding Radicalization Pathways: A Framework for Assessing Diversity in YouTube Recommendation Systems". In: *Social Network Analysis and Mining* 14.1 (2024), pp. 1–11.

[15] John M Betts and Ana-Maria Bliuc. "The Effect of Influencers on Societal Polarization". In: *2022 Winter Simulation Conference (WSC)*. IEEE. 2022, pp. 370–381.

[16] Nikita Bhalla, Adam Lechowicz, and Cameron Musco. "Local Edge Dynamics and Opinion Polarization". In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 2023, pp. 6–14.

[17] María J Blanca et al. "Skewness and Kurtosis in Real Data Samples". In: *Methodology* (2013).

[18] Vincent D. Blondel et al. "Fast Unfolding of Communities in Large Networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008.

[19] Fernando H Calderón et al. “Content-Based Echo Chamber Detection on Social Media Platforms”. In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2019, pp. 597–600.

[20] Campaign for Accountability. *YouTube Hosts Hundreds of Militia Videos Fueling Extremist Movement, Ukraine Disinformation*. https://campaignforaccountability.org/ttp-report-youtube-hosts-hundreds-of-militia-videos-fueling-extremist-movement-ukraine-disinformation. Accessed: 2025-08-05. 2022.

[21] Ahmet Celikoglu and Ugur Tirnakli. “Skewness and Kurtosis Analysis for Non-Gaussian Distributions”. In: *Physica A: Statistical Mechanics and its Applications* 499 (2018), pp. 325–334.

[22] Nishan Chatterjee. *Multilingual Persuasion Detection from Text*. Accessed: 2025-05-12. 2024. URL: https://huggingface.co/nishan-chatterjee/multilingual-persuasion-detection-from-text.

[23] Ling Chen et al. “Random Walk-Based Algorithm for Distance-Aware Influence Maximization on Multiple Query Locations”. In: *Knowledge-Based Systems* 249 (2022), p. 108820.

[24] Sijing Chen and Lu Xiao. “Predicting and Characterising Persuasion Strategies in Misinformation Content over Social Media Based on the Multi-Label Classification Approach”. In: *Journal of Information Science* (2023), p. 01655515231169949.

[25] Matteo Cinelli et al. “The Echo Chamber Effect on Social Media”. In: *Proceedings of the National Academy of Sciences* 118.9 (2021), e2023301118.

[26] Federico Cinus et al. “The Effect of People Recommenders on Echo Chambers and Polarization”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 90–101.

[27] Corinna Coupette, Stefan Neumann, and Aristides Gionis. “Reducing Exposure to Harmful Content via Graph Rewiring”. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023, pp. 323–334.

[28] James Crowley. “Full List of Celebrities Endorsing Donald Trump: Kid Rock, Mel Gibson, and More”. In: *Newsweek* (Nov. 2024). Accessed: 2025-08-05. URL: https://www.newsweek.com/celebrities-endorsing-donald-trump-kid-rock-brett-favre-1977249.

[29] Anthony Cuthbertson. *Facebook Membership of Anti-Mask Groups Shoots Up Nearly 2000% Since August*. Accessed: 2025-07-28. 2020. URL: https://www.independent.co.uk/tech/facebook-anti-mask-groups-covid-19-coronavirus-august-b1711234.html.

[30] Elizabeth M Daly, Werner Geyer, and David R Millen. “The Network Effects of Recommending Social Connections”. In: *Proceedings of the Fourth ACM Conference on Recommender Systems*. 2010, pp. 301–304.

[31] Elizabeth M Daly, Werner Geyerand, and David R Millen. “The Network Effects of Recommending Social Connections”. In: *Proceedings of the Fourth ACM Conference on Recommender Systems*. 2010, pp. 301–304.

[32] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[33] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186.

[34] Niccolò Di Marco, Matteo Cinelli, and Walter Quattrociocchi. “Infodemics on YouTube: Reliability of Content and Echo Chambers on COVID-19”. In: *arXiv preprint arXiv:2106.08684* (2021).

[35] Fernando Diaz-Diaz, Maxi San Miguel, and Sandro Meloni. “Echo Chambers and Information Transmission Biases in Homophilic and Heterophilic Networks”. In: *Scientific Reports* 12.1 (2022), p. 9350.

[36] United Nations Office on Drugs and Crime. *The Role of Social Media in Terrorism and Radicalization*. Accessed: YYYY-MM-DD. 2021. URL: https://www.unodc.org.

[37] Sabine A. Einwiller and Sora Kim. “How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication: An Analysis of Policies and Their Implementation”. In: *Policy & Internet* 12.2 (2020), pp. 184–206.

[38] Aiman El Asam and Muthanna Samara. “Cyberbullying and the Law: A Review of Psychological and Legal Challenges”. In: *Computers in Human Behavior* 65 (2016), pp. 127–141.

[39] Francesco Fabbri et al. “Rewiring What-to-Watch-Next Recommendations to Reduce Radicalization Pathways”. In: *Proceedings of the ACM Web Conference 2022*. 2022, pp. 2719–2728.

[40] Tasja-Selina Fischer, Castulus Kolo, and Cornelia Mothes. “Political Influencers on YouTube: Business Strategies and Content Characteristics”. In: *Media and Communication* 10.1 (2022), pp. 259–271.

[41] James Flamino et al. “Political Polarization of News Media and Influencers on Twitter in the 2016 and 2020 US Presidential Elections”. In: *Nature Human Behaviour* 7.6 (2023), pp. 904–916.

[42] Noah E. Friedkin and Eugene C. Johnsen. “Social Influence and Opinions”. In: *Journal of Mathematical Sociology* 15.3-4 (1990), pp. 193–206.

[43] Chongming Gao et al. “CIRS: Bursting Filter Bubbles by Counterfactual Interactive Recommender System”. In: *ACM Transactions on Information Systems* 42.1 (2023), pp. 1–27.

[44] Lanlin Gao et al. “Discriminating Topical Influencers Based on the User Relative Emotion”. In: *IEEE Access* 7 (2019), pp. 100120–100130.

[45] Yichang Gao, Fengming Liu, and Lei Gao. “Echo Chamber Effects on Short Video Platforms”. In: *Scientific Reports* 13.1 (2023), p. 6282.

[46] Kiran Garimella et al. “Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship”. In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 913–922.

[47] Kiran Garimella et al. “Quantifying Controversy on Social Media”. In: *ACM Transactions on Social Computing* 1.1 (2018), pp. 1–27.

[48] Shuang Geng et al. “Accuracy-Diversity Optimization in Personalized Recommender System via Trajectory Reinforcement Based Bacterial Colony Optimization”. In: *Information Processing & Management* 60.2 (2023), p. 103205.

[49] Vahid Ghafouri et al. “NLP-Driven Approaches to Measuring Online Polarization and Radicalization”. PhD thesis. Universidad Carlos III de Madrid, Spain, 2025.

[50] Vahid Ghafouri et al. “Transformer-Based Quantification of the Echo Chamber Effect in Online Communities”. In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW2 (2024), pp. 1–27.

[51] Tarleton Gillespie. “Do Not Recommend? Reduction as a Form of Content Moderation”. In: *Social Media + Society* 8.3 (2022), p. 20563051221117552.

[52] Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. “Social Media, Sentiment and Public Opinions: Evidence from #Brexit and #USElection”. In: *European Economic Review* 136 (2021), p. 103772.

[53] Kamile Grusauskaite et al. “Debating (in) Echo Chambers: How Culture Shapes Communication in Conspiracy Theory Networks on YouTube”. In: *New Media & Society* 26.12 (2024), pp. 7037–7057.

[54] Chang Guo et al. “Heterogeneous Network Influence Maximization Algorithm Based on Multi-Scale Propagation Strength and Repulsive Force of Propagation Field”. In: *Knowledge-Based Systems* 291 (2024), p. 111580.

[55] Shahrzad Haddadan et al. “Repbublik: Reducing Polarized Bubble Radius with Link Insertions”. In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 2021, pp. 139–147.

[56] Will Hamilton, Zhitao Ying, and Jure Leskovec. “Inductive Representation Learning on Large Graphs”. In: *Advances in Neural Information Processing Systems* 30 (2017).

[57] Muhammad Haroon et al. “Auditing YouTube’s Recommendation System for Ideologically Congenial, Extreme, and Problematic Recommendations”. In: *Proceedings of the National Academy of Sciences* 120.50 (2023), e2213020120.

[58] Muhammad Haroon et al. “YouTube, the Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations”. In: *arXiv preprint arXiv:2203.10666* (2022).

[59] Muhammad Haroon et al. “YouTube, the Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations”. In: *arXiv preprint arXiv:2203.10666* (2022).

[60] David Hartmann et al. “A Systematic Review of Echo Chamber Research: Comparative Analysis of Conceptualizations, Operationalizations, and Varying Outcomes”. In: *Journal of Computational Social Science* 8.2 (2025), p. 52.

[61] Jochen Hartmann. *Emotion English DistilRoBERTa-Base*. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/. 2022.

[62] Luzie Helfmann et al. “Modelling Opinion Dynamics Under the Impact of Influencer and Media Strategies”. In: *Scientific Reports* 13.1 (2023), p. 19375. DOI: 10.1038/s41598-023-46187-9. URL: https://doi.org/10.1038/s41598-023-46187-9.

[63] Itai Himelboim, Stephen McCreery, and Marc Smith. “Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter”. In: *Journal of Computer-Mediated Communication* 18.2 (2013), pp. 154–174.

[64] Man Chun Hui. *US Election 2020 Tweets*. Accessed: 2025-01-16. 2020. URL: https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets.

[65] Paola Impicciché and Marco Viviani. “Comparing Echo Chamber Detection Metrics: A Cross-Modeling and Cross-Platform Analysis of Twitter and Reddit”. In: *ACM Transactions on the Web* (2024).

[66] Dietmar Jannach et al. *Recommender Systems: An Introduction*. Cambridge University Press, 2010.

[67] Brent Kitchens, Steven L. Johnson, and Peter Gray. “Understanding Echo Chambers and Filter Bubbles: The Impact of Social Media on Diversification and Partisan Shifts in News Consumption”. In: *MIS Quarterly* 44.4 (2020).

[68] Ofra Klein. “Anti-Immigrant Rhetoric of Populist Radical Right Leaders on Social Media Platforms”. In: *Communications* 49.3 (2024), pp. 400–420.

[69] Nane Kratzke. "How to Find Orchestrated Trolls? A Case Study on Identifying Polarized Twitter Echo Chambers". In: *Computers* 12.3 (2023), p. 57.

[70] Lucio La Cava, Luca Aiello, and Andrea Tagarelli. "Drivers of Social Influence in the Twitter Migration to Mastodon". In: *Scientific Reports* 13 (Dec. 2023). DOI: 10.1038/s41598-023-48200-7.

[71] Nicolas Lanzetti, Florian Dörfler, and Nicolò Pagan. "The Impact of Recommendation Systems on Opinion Dynamics: Microscopic versus Macroscopic Effects". In: *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE. 2023, pp. 4824–4829.

[72] David MJ Lazer et al. "The Science of Fake News". In: *Science* 359.6380 (2018), pp. 1094–1096.

[73] Mark Ledwich and Anna Zaitsev. "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization". In: *arXiv preprint arXiv:1912.11211* (2019).

[74] Zhenyang Li et al. "Breaking Filter Bubble: A Reinforcement Learning Framework of Controllable Recommender System". In: *Proceedings of the ACM Web Conference 2023*. 2023, pp. 4041–4049.

[75] Yinhan Liu. "Roberta: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[76] Daniel Loureiro et al. "TimeLMs: Diachronic Language Models from Twitter". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 251–260. DOI: 10.18653/v1/2022.acl-demo.25. URL: https://aclanthology.org/2022.acl-demo.25.

[77] Md Ishtyaq Mahmud, Muntasir Mamun, and Ahmed Abdelgawad. "A Deep Analysis of Textual Features Based Cyberbullying Detection Using Machine Learning". In: *2022 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*. IEEE. 2022, pp. 166–170.

[78] Katerina Eva Matsa and Elisa Shearer. "News Use Across Social Media Platforms 2018". In: *Pew Research Center* 10 (2018).

[79] James B McDonald, Jeff Sorensen, and Patrick A Turley. "Skewness and Kurtosis Properties of Income Distribution Models". In: *Review of Income and Wealth* 59.2 (2013), pp. 360–374.

[80] Meta. *Community Standards Enforcement Report Q1 2020*. https://about.fb.com/news/2020/05/combating-hate-and-dangerous-organizations. Accessed: 2025-08-05. May 2020.

[81] Marco Minici et al. "Cascade-Based Echo Chamber Detection". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022, pp. 1511–1520.

[82] Volodymyr Mnih et al. "Playing Atari with Deep Reinforcement Learning". In: *arXiv preprint arXiv:1312.5602* (2013).

[83] Virginia Morini, Laura Pollacci, and Giulio Rossetti. "Toward a Standard Approach for Echo Chamber Detection: Reddit Case Study". In: *Applied Sciences* 11.12 (2021), p. 5390.

[84] Pau Muñoz et al. "Quantifying Polarization in Online Political Discourse". In: *EPJ Data Science* 13.1 (2024), p. 39.

[85] Khalid T Mursi et al. "Detecting Islamic Radicalism Arabic Tweets Using Natural Language Processing". In: *IEEE Access* 10 (2022), pp. 72526–72534.

[86] Luisa Muth and Christina Peter. "Social Media Influencers' Role in Shaping Political Opinions and Actions of Young Audiences". In: *Media and Communication* 11.3 (2023), pp. 164–174.

[87] Andrew Mvd. *Cyberbullying Classification*. https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification. Accessed: 2024-11-18. 2022.

[88] M Nisha and J Jebathangam. "Detection and Classification of Cyberbullying in Social Media Using Text Mining". In: *2022 6th International Conference on Electronics, Communication and Aerospace Technology*. IEEE. 2022, pp. 856–861.

[89] Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adalí. "NELA-GT-2018: A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 13. 2019, pp. 630–638.

[90] Observatory on Social Media (OSoMe). *Tracking Public Opinion About Unsupported Narratives in the 2020 Presidential Election – Wave 6 Report*. Tech. rep. Data Collected Nov 18–27, 2020. Observatory on Social Media, Indiana University, Nov. 2020. URL: https://osome.iu.edu/research/white-papers/Tracking%20Public%20Opinion%20Wave%206.pdf.

[91] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin UK, 2011.

[92] Chaïm Perelman. *The New Rhetoric and the Humanities: Essays on Rhetoric and its Applications*. Vol. 140. Synthese Library. With an introduction by Harold Zyskind. Dordrecht: D. Reidel Publishing Company, 1979. ISBN: 902771018X.

[93] Nicola Perra and Luis EC Rocha. "Modelling Opinion Dynamics in the Age of Algorithmic Personalisation". In: *Scientific Reports* 9.1 (2019), p. 7261.

[94] Lorenzo Porcaro, Carlos Castillo, and Emilia Gómez Gutiérrez. "Diversity by Design in Music Recommender Systems". In: *Transactions of the International Society for Music Information Retrieval* 4.1 (2021).

[95] Priori Data. *Global Social Media Usage Statistics (2018–2025)*. Accessed: 2025-05-23. 2025. URL: `https://prioridata.com/data/social-media-usage/`.

[96] Pearl Pu, Li Chen, and Rong Hu. "Evaluating Recommender Systems from the User's Perspective: Survey of the State of the Art". In: *User Modeling and User-Adapted Interaction* 22 (2012), pp. 317–355.

[97] Inzamam Rahaman and Patrick Hosein. "A Model for Optimizing Article Recommendation for Reducing Polarization". In: *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2021, pp. 107–110.

[98] Manoel Horta Ribeiro et al. "Auditing Radicalization Pathways on YouTube". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 131–141.

[99] Manoel Horta Ribeiro et al. "Auditing Radicalization Pathways on YouTube". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 131–141.

[100] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?" Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

[101] Magdalena Riedl et al. “The Rise of Political Influencers—Perspectives on a Trend Towards Meaningful Content”. In: *Frontiers in Communication* 6 (2021), p. 752656.

[102] Daniel Röchert et al. “The Networked Context of COVID-19 Misinformation: Informational Homogeneity on YouTube at the Beginning of the Pandemic”. In: *Online Social Networks and Media* 26 (2021), p. 100164.

[103] Hiteshi Saini et al. “Enhancing Cyberbullying Detection: A Comparative Study of Ensemble CNN–SVM and BERT Models”. In: *Social Network Analysis and Mining* 14.1 (2023), p. 1.

[104] Fernando P Santos, Yphtach Lelkes, and Simon A Levin. “Link Recommendation Algorithms and Dynamics of Polarization in Online Social Networks”. In: *Proceedings of the National Academy of Sciences* 118.50 (2021), e2102141118.

[105] Desirée Schmuck et al. “Politics—Simply Explained? How Influencers Affect Youth’s Perceived Simplification of Politics, Political Cynicism, and Political Interest”. In: *The International Journal of Press/Politics* 27.3 (2022), pp. 738–762.

[106] Jack Selzer. “Rhetorical Analysis: Understanding How Texts Persuade Readers”. In: *What Writing Does and How It Does It*. Routledge, 2003, pp. 285–314.

[107] Donghee Shin and Kulsawasd Jitkajornwanich. “How Algorithms Promote Self-Radicalization: Audit of TikTok’s Algorithm Using a Reverse Engineering Method”. In: *Social Science Computer Review* 42.4 (2024), pp. 1020–1040.

[108] Emily Singer. “Taylor Swift Endorses Kamala Harris: ‘She Is a Steady-Handed, Gifted Leader’”. In: *The Pennsylvania Independent* (Sept. 2024). Accessed: 2025-08-05. URL: https://pennsylvaniaindependent.com/politics/taylor-swift-endorses-kamala-harris-she-is-a-steady-handed-gifted-leader/.

[109] Felipe Bonow Soares, Raquel Recuero, and Gabriela Zago. “Influencers in Polarized Political Networks on Twitter”. In: *Proceedings of the 9th International Conference on Social Media and Society*. 2018, pp. 168–177.

[110] Ivan Srba et al. “Auditing YouTube’s Recommendation Algorithm for Misinformation Filter Bubbles”. In: *ACM Transactions on Recommender Systems* 1.1 (2023), pp. 1–33.

[111] Statista. *X (Formerly Twitter): Enforcement of Violence and Hate Content 2024*. `https://www.statista.com/statistics/1497510/x-twitter-enforcement-violence-and-hate`. Accessed: 2025-08-05. 2024.

[112] Jessica Su, Aneesh Sharma, and Sharad Goel. “The Effect of Recommendations on Network Structure”. In: *Proceedings of the 25th International Conference on World Wide Web*. 2016, pp. 1157–1167.

[113] Haoxin Sun and Zhongzhi Zhang. “Opinion Optimization in Directed Social Networks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 4. 2023, pp. 4623–4632.

[114] Mingfei Sun, Xiaoyue Ma, and Yudi Huo. “Does Social Media Users’ Interaction Influence the Formation of Echo Chambers? Social Network Analysis Based on Vaccine Video Comments on YouTube”. In: *International Journal of Environmental Research and Public Health* 19.23 (2022), p. 15869.

[115] Ye Tian and Long Wang. “Opinion Dynamics in Social Networks with Stubborn Agents: An Issue-Based Perspective”. In: *Automatica* 96 (2018), pp. 213–223.

[116] Antonela Tommasel and Filippo Menczer. “Do Recommender Systems Make Social Media More Susceptible to Misinformation Spreaders?” In: *Proceedings of the 16th ACM Conference on Recommender Systems*. 2022, pp. 550–555.

[117] Muhammad Umer et al. “Cyberbullying Detection Using PCA Extracted GLOVE Features and RoBERTaNet Transformer Learning Model”. In: *IEEE Transactions on Computational Social Systems* (2024).

[118] Oswelled Ureke. “Politics at Play: TikTok and Digital Persuasion in Zimbabwe’s 2023 General Elections”. In: *Africa Spectrum* 59.2 (2024), pp. 254–278.

[119] Jay J Van Bavel et al. “How Social Media Shapes Polarization”. In: *Trends in Cognitive Sciences* 25.11 (2021), pp. 913–916.

[120] Stefaan Vercoutere et al. “Improving Selection Diversity Using Hybrid Graph-Based News Recommenders”. In: *User Modeling and User-Adapted Interaction* 34.4 (2024), pp. 955–993.

[121] Giacomo Villa, Gabriella Pasi, and Marco Viviani. "Echo Chamber Detection and Analysis: A Topology- and Content-Based Approach in the COVID-19 Scenario". In: *Social Network Analysis and Mining* 11.1 (2021), p. 78.

[122] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The Spread of True and False News Online". In: *Science* 359.6380 (2018), pp. 1146–1151.

[123] Dandan Wang, Yadong Zhou, and Feicheng Ma. "Opinion Leaders and Structural Hole Spanners Influencing Echo Chambers in Discussions About COVID-19 Vaccines on Social Media in China: Network Analysis". In: *Journal of Medical Internet Research* 24.11 (2022), e40701.

[124] Jason Wang, Kaiqun Fu, and Chang-Tien Lu. "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection". In: *2020 IEEE International Conference on Big Data (Big Data)*. IEEE. 2020, pp. 1699–1708.

[125] Mengyan Wang et al. "Nudging Towards Responsible Recommendations: A Graph-Based Approach to Mitigate Belief Filter Bubbles". In: *IEEE Transactions on Artificial Intelligence* (2024).

[126] Joe Whittaker et al. "Recommender Systems and the Amplification of Extremist Content". In: *Internet Policy Review* 10.2 (2021).

[127] Wired. *X's First Transparency Report Under Musk Shows 224 Million User Reports*. https://www.wired.com/story/x-twitter-first-transparency-report-since-elon-musks-takeover-is-finally-here. Accessed: 2025-08-05. 2024.

[128] Magdalena E Wojcieszak and Diana C Mutz. "Online Groups and Political Discourse: Do Online Discussion Spaces Facilitate Exposure to Political Disagreement?" In: *Journal of Communication* 59.1 (2009), pp. 40–56.

[129] Youping Xu. "The Invisible Aggressive Fist: Features of Cyberbullying Language in China". In: *International Journal for the Semiotics of Law-Revue Internationale de Sémiotique Juridique* 34.4 (2021), pp. 1041–1064.

[130] Rocío Zamora-Medina, Andrius Suminas, and Shahira S. Fahmy. "Securing the Youth Vote: A Comparative Analysis of Digital Persuasion on TikTok among Political Actors". In: *Media and Communication* 11.2 (2023), pp. 218–231.

[131] John Zarocostas. "How to Fight an Infodemic". In: *The Lancet* 395.10225 (2020), p. 676.

[132] Chunkai Zhang, Guoqing Li, and Hanyu Zhang. "Multi-View Contrastive Learning with Virtual Social Group Influence for Social Recommendation". In: *Knowledge-Based Systems* 294 (2024), p. 111751.

[133] Kun Zhang et al. "Towards Identifying Influential Nodes in Complex Networks Using Semi-Local Centrality Metrics". In: *Journal of King Saud University-Computer and Information Sciences* 35.10 (2023), p. 101798.

[134] Xiaohang Zhang et al. "Identifying Influential Nodes in Complex Networks with Community Structure". In: *Knowledge-Based Systems* 42 (2013), pp. 74–84.

[135] Zhili Zhao et al. "Ranking Influential Spreaders Based on Both Node k-Shell and Structural Hole". In: *Knowledge-Based Systems* 260 (2023), p. 110163.

[136] Qinyue Zhou and Zhibin Wu. "Multidimensional Friedkin-Johnsen Model with Increasing Stubbornness in Social Networks". In: *Information Sciences* 600 (2022), pp. 170–188.

[137] Xiaotian Zhou et al. "Friedkin-Johnsen Model for Opinion Dynamics on Signed Graphs". In: *IEEE Transactions on Knowledge and Data Engineering* (2024).