Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REVEALING BIAS DIFFUSION IN SOCIAL MEDIA THROUGH STRUCTURAL AND SEMANTIC MEASURES

ELIZABETH LEAH GEORGE[1] , SUBASHINI PARTHASARATHY [2]

[1]Research Scholar, Department of Computer Science,

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

[2]Professor, Centre for Machine Learning and Intelligence (CMLI), Department of Computer Science,

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India Country

E-mail: [1]18phcsp003@avinuity.ac.in, [2]subashini_cs@avinuty.ac.in

ABSTRACT

Social media has become a powerful force shaping public narratives, particularly on sensitive topics. However, these platforms also allow biased or unfair language to spread quickly. This study examines the diffusion of biased language on violent crime through Twitter discourse (using the hashtag #Murder). Understanding such bias diffusion is crucial because it can skew public perception and potentially impact judicial outcomes. Our goal is to measure and mitigate bias by combining advanced language models and network analysis. We propose a framework that integrates semantic language processing and structural topology analysis. A transformer-based model (T5) extracts semantic features from tweets, and to test bias mitigation, we apply text style transfer to rewrite content into neutral, positive, or negative forms. We build a two-layer multiplex concept network: one layer for general (unbiased) concepts from term co-occurrence and another for bias-related concepts identified via a curated lexicon of sensitive terms links between layers capture where neutral and bias-laden concepts co-occur, revealing bias infiltration pathways. We quantify structural and semantic similarity using Jaccard similarity, graph edit distance, cosine similarity, DeltaCon, and NetSimile metrics, and we measure bias propagation using a bias infiltration index. Bias-related terms are fully integrated into the concept network in the original tweet corpus (Jaccard node overlap = 1.0, Weighted Jaccard = 1.0), indicating that biased language co-occurs with all key topics. After tweets are rewritten by sentiment, this overlap decreases: the overall node overlap drops to ~0.90, and the weighted overlap declines most in neutral rewrites (to ~0.76, versus ~0.88 in positive/negative cases). Positive tweets embed bias most strongly, whereas neutral rewrites retain a high degree of subtle bias (as indicated by the Bias Infiltration Index). Statistical tests (paired t-tests, $p > 0.49$) confirm that the reductions in explicit bias are not significant, underscoring the bias's resilience. Overall, our findings suggest that while sentiment-based rewriting can partly disrupt bias pathways, biased concepts remain deeply woven into social media discourse.

Keywords: *Social Media Discourse, Bias Detection, Semantic Embeddings, Text Style Transfer, Concept Maps*

1. INTRODUCTION

Social media platforms, such as Twitter, enable instant information and opinion sharing, making them powerful public discourse drivers. However, these platforms also allow the rapid spread of biased or influenced language, which can distort public understanding and intensify social divisions. This effect is especially pronounced around sensitive topics: social media discussions of high-profile crimes often become sharply polarized.

Research into the murder of Brazilian councilwoman Marielle Franco [1] revealed that online sentiment experienced a considerable change after the event, whereas an examination of the Camilo Catrillanca case in Chile uncovered echo chambers that perpetuated biased narratives.[2] Research with mock juries found that exposure to negative social media comments can influence verdict decisions. Simultaneously, social media can fuel positive change: the global movements sparked

by the killing of George Floyd illustrate how online discourse can mobilize calls for justice and reform.[3] These examples highlight the dual role of social media in amplifying prejudice and enabling activism, [4] underscoring the real-world significance of addressing bias in online discourse.

The bias problem needs urgent attention: Bias in social media discourse is not merely an academic curiosity: it has measurable downstream consequences for public opinion formation, civic trust, and potentially legal outcomes. Online bias can accelerate the formation of echo-chambers, amplify inaccurate or decontextualized narratives, and produce rapid shifts in sentiment that influence policymaking and collective action. For sensitive topics such as violent crime, biased language can shape framing around victims and suspects, influence bystander perception, and leak into offline decision-making (for example, juror attitudes or press coverage). Because social platforms are now major channels for news and civic discussion, even subtle, distributed bias can create persistent structural pathways that magnify marginal signals into system-level effects. Addressing bias diffusion how biased wording becomes structurally integrated into topic networks therefore has clear societal importance: reducing unfair framing may lower misinformation spill over, mitigate stigmatization of groups, and improve the fairness of automated downstream systems (e.g., trend detection, recommendation). This study, therefore, prioritizes not only the detection of lexical bias but the measurement of its structural embedding within concept networks and the exploration of interventions that alter those structural pathways.

Despite advances in natural language processing and network analysis, few studies have systematically integrated these approaches to track how bias evolves and spreads in online communities [5]. Previous studies have often analyzed textual bias or identified polarized user groups, but a unified framework to capture both semantic content and structural dynamics of bias diffusion remains underexplored. This study addresses this gap by applying multilayer network analysis together with bias-sensitive language models to a corpus of murder-related tweets. Concept networks are constructed that separate neutral co-occurrences from bias-laden terms and examine the links between them. This framework reveals how polarized narratives cluster and diffuse, shaping public discourse on crime-related topics.

This study focuses on textual analysis of Twitter conversations about murder and concept-level relationships within those tweets. The scope and hypothesis of the study are as follows:

- Scope: We analyze English-language tweets tagged #Murder, constructing multiplex concept networks that capture relationships between general and bias-related terms. We apply sentiment-based style transfer to create neutral, positive, and negative variants of each tweet for comparison.
- Limitations: We do not analyze non-textual content (such as images or videos) or tweets from other platforms. Our analysis is limited to linguistic bias as defined by our curated lexicon and does not attempt to detect all ideological bias or misinformation. Integrating bias-aware natural language processing systems with network analysis may uncover concealed patterns of bias propagation across various sentiment levels, and sentiment-driven rewriting could potentially mitigate the dissemination of biased language.

1.1 Research Hypothesis

Hypotheses. On the theoretical premise that biased lexical items can become structurally embedded in discourse networks, we test the following hypotheses:

- In H1(Structural embedding): bias-related terms are disproportionately embedded in the core structure of the concept network for #Murder tweets, with bias terms displaying high centrality and bridging roles between topical groups.
- H2 (Mitigation by rewriting): Sentiment-based text style transfer (neutral, positive, negative rewrites) reduces the structural integration of bias as measured by the Bias Infiltration Index and weighted similarity metrics (Weighted Jaccard, DeltaCon), although it may not substantially reduce the raw lexical counts of bias terms.
- H3 (Style-specific effects): Different rewrite styles alter network embedding differently neutral rewrites are expected to reduce weighted co-occurrence strength more than positive or negative rewrites, but some subtle bias cues may persist.

The remainder of this research is organized as follows: **Section 2** presents the Related Work focused on concept map construction, bias detection models, network similarity measures, and knowledge graph analytics in social media. **Section 3** describes the Proposed Methodology in detail, outlining the research design and implementation. Emperical pipeline showcases

prrocessing to multiplex network building.The application of structural and semantic similarity metrics. **Section 4** elaborates on the Experimental Setup, including System configuration, model implementation, classification and style transfer of tweets, with visualization techniques. **Section 5** reports and discusses the Results, highlighting quantitative metrics , Statistical significance tests to analyse network similarity analysis, bias infiltration metrics, with visualization outputs demonstrating how bias propagates and transforms across different sentiment layers. Lastly, **Section 6** summarizes the key conclusions of the study and proposes directions for Future Research, including recommendations for advancing automated bias detection and diffusion analysis in digital social networks.

2. RELATED WORK

Social media platforms create complex information ecosystems where bias can propagate through both linguistic content and network structures, making traditional single-modal approaches [6] insufficient for capturing their dynamics. To address this, several computational frameworks have been proposed to track and interpret bias diffusion. For instance, the BiasWatch System [7] adopts a network-based approach to identify opinion leaders and trace bias themes, achieving a 20% improvement in bias estimation accuracy and a 26% increase in user recommendation precision. Similarly, Concept Relevance Propagation (CRP) [8] integrates local and global interpretability to provide deeper insights into the location and nature of bias within models, enhancing transparency in bias detection. Complementing these methods, the DivDist Framework [9] grounds bias measurement in social science principles and validates its constructs through rigorous empirical testing, offering a more trustworthy approach to quantifying social.

Beyond system-level approaches, concept maps and semantic networks have been widely applied to capture how ideas and terms connect in online conversations.[10] When combined with graph-based methods, concept maps reveal both the structural pathways and semantic overlaps through which bias infiltrates discourse. This is particularly relevant to politically and socially sensitive discussions, such as those surrounding murder cases on Twitter, where biased narratives often reinforce echo chambers and polarization [11].

Another emerging direction for bias mitigation is Text Style Transfer (TST), which alters the stylistic properties of text while preserving semantic meaning. TST has been shown to reduce bias in domains such as job advertisements, where gendered language discourages applications from under represented groups [13]. Techniques combining latent content encoding with explicit keyword replacement have improved both style transfer quality and content fidelity. More recently, diffusion-based probabilistic models have advanced the field by enabling fine-grained control of stylistic transformations.[12] These models outperform traditional approaches even in low-resource settings and can operate without reliance on external knowledge sources, making them particularly suitable for social media contexts where textual data is unstructured and biased [14]. However, researchers note that TST may also introduce unintended biases, underscoring the importance of integrating these methods with network-based bias analysis to ensure both effectiveness and interpretability.

Taken together, prior work highlights three complementary perspectives on bias in digital discourse: network-based frameworks that trace the spread of bias, concept maps that visualize semantic and structural connections [15]-[18], and text style transfer techniques, that mitigate bias at the linguistic level. Yet, despite these advances, relatively few studies combine these strands into an integrated framework capable of both detecting and explaining how bias emerges, diffuses, and transforms within social media conversations. Concept maps serve as powerful tools for visualizing the semantic and structural interrelationships [19] between ideas and biased themes, enabling researchers to map both explicit and implicit associations that shape online discourse. Meanwhile, advances in text style transfer and transformer-based methods allow for the modification or neutralization of biased language, offering avenues to not just detect but actively mitigate and reshape problematic discourse.[20-23] Despite these technological and conceptual advances, there are few studies that integrate these strands into a unified analytical framework, one capable of both detecting and explaining how bias emerges, spreads, and transforms in the evolving landscape of social media conversation. [24-26] Most contemporary efforts either focus on network structure, semantic mapping, or linguistic bias mitigation [28] in isolation, indicating a gap in fully integrated approaches that holistically account for the complexity of bias propagation and transformation in digital spaces

This study addresses this gap by examining the infiltration of bias in Twitter discourse surrounding the hashtag #Murder, constructing a multilayer network that combines semantic concept maps with bias-driven subgraphs, and applying both structural and semantic measures to quantify bias diffusion. Integrating concept maps help us see how different words and ideas are connected in social media conversations. When used with network analysis, they can show not only what people are talking about but also how bias moves through these conversations.

3. PROPOSED METHODOLOGY

3.1 Research design and implementation steps

This study follows a reproducible, empirical pipeline that combines natural language processing and network science. The research design is quantitative and comparative, using original tweets and sentiment-altered rewrites to test how language changes affect structural bias. The major steps are:

- ***Data collection and filtering*:** Compile the #Murder tweet corpus (English-language tweets only), remove duplicates, non-English tweets, and bot-like accounts, and retain tweets with substantive text. (Curated Dataset collected: 32,000 tweets)
- ***Preprocessing and tokenization:*** Tokenize, normalize, remove stopwords, and preserve relevant identity-bearing tokens by applying named-entity recognition.
- ***Curation and validation of bias lexicons:*** Build and validate a curated lexicon of socially/politically sensitive terms (bias candidates) via bias detection model that checks across the corpus; creates document lexicon and inclusion criteria.
- ***Semantic feature extraction:*** Generate sentence and token embeddings using transformer models (e.g., all-MiniLM-L6-v2 for embeddings; T5-based encoders for augmented features) [27],[29] and extract thematic keywords and co-occurrences.
- ***Text style transfer:*** Produce neutral, positive, and negative rewrites using controlled paraphrasing (parrot_paraphraser_on_T5 [30] or equivalent). Semantic fidelity checks (e.g., cosine similarity thresholds) are maintained during rewrites to ensure that topicality is preserved.
- ***To construct a concept map:*** Co-occurrence graphs using tweet-level co-occurrence approach are generated to create the concept graph and a distinct bias subgraph based on matches with a bias lexicon; node and edge weights are then determined based on frequency or embedding similarity.
- ***Multiplex network assembly*:** The concept graph and bias subgraph are combined into a two-layer multiplex network with interlayer semantic alignment links (i.e., node alignment via embeddings or lexical matching).
- ***Metric computation and statistical analysis:*** structural and semantic similarity metrics (Jaccard, weighted Jaccard, Graph Edit Distance, Cosine adjacency, DeltaCon, NetSimile) and a composite Bias Infiltration Index (BII). Pair statistical tests (e.g., paired t-tests with effect sizes and confidence intervals) were used to compare the original and rewritten networks.
- ***Visualization and robustness checks:*** Produce topology visualizations and heatmaps and perform sensitivity checks (e.g., variance co-occurrence window, lexicon thresholds, and embedding similarity cutoffs).

3.2 Empirical Pipeline of Research

The research employs a multi-stage methodology to analyze and quantify bias propagation in Twitter discourse. a systematic workflow to graph and evaluate knowledge diffusion on social media using automated concept maps and network similarity measures is a unique approach for #murder related tweets that could include bias related study. This study was carried out to identify user bias and how much of truthful data is revealed owing to bias in social media. The text style transfer when applied to tweets can alters the stylistic properties of text while preserving semantic meaning. TST has been shown to reduce bias in tweets.

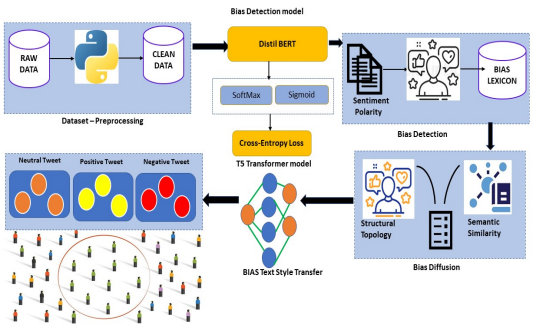


Figure 1. Proposed Framework With Augmented Feature Extraction

The proposed methodology in Figure 1. combines automated natural language processing and advanced network analysis to objectively graph knowledge and measure bias propagation within social media tweet corpora. The process commences with dataset pre-processing, where raw Twitter data are cleaned tokenizing text, removing noise, and extracting core keywords yielding a high-quality set of tweet content suitable for computational analysis. The workflow, depicted in the provided diagram, consists of the following major phases:

3.2.1 Dataset Pre-processing

Raw Twitter data is collected and subjected to a comprehensive pre-processing pipeline. This stage includes cleaning the data to remove noise, irrelevant information, and inconsistencies, ensuring that only meaningful and high-quality tweets are used for analysis. The sentiment approach of each tweet is analysed, and only meaningful tweets are considered. The analysis of the tweets revealed more politically and socially sensitive tweets, which encouraged further study and bias analysis.

3.2.2 Bias Detection Models

Once pre-processed, the clean tweets are analyzed using bias detection models. These models apply a curated bias lexicon to identify and tag biased language, enabling systematic detection of potentially biased terms and phrases within the dataset. The bias detection phase leverages both machine learning models and curated bias lexicons to systematically identify and label bias expressions within each tweet.

3.2.3 Structural Topology and Semantic Similarity Analysis

For each tweet category, the structural topology (network structure) and semantic similarity (lexical relatedness) are analyzed. These quantitative measures reveal how biased language shapes the connections between concepts and the semantic content of discussions. With both original and sentiment-altered networks created, a suite of structural and semantic network similarity measures is applied. Metrics such as: **Jaccard similarity** (for nodes and edges) compares overlap in concepts and relationships, Cosine similarity: Quantifies vectorized similarity between two concept maps. **Graph Edit Distance**: Measures the minimum number of changes (additions or deletions of nodes/edges) required to transform one graph into another.**Weighted Jaccard**: Quantifies the overlap between two graphs by considering the sum of the minimum edge weights divided by the sum of the maximum edge weights for shared connections.**DeltaCon**: Assesses the similarity between two graphs based on how easily information can diffuse through both networks, capturing fine-grained structural differences.
NetSimile: Compares graphs by extracting and statistically comparing node-level network features to characterize overall graph similarity.
Bias Infiltration Index (BII): A composite score reflecting the extent to which bias-related language permeates and structurally integrates within a larger concept or discourse network.

All the above metrics are computed to quantitatively capture the degree of overlap, change, and structural or semantic embedding of biased language between networks. A composite Bias Infiltration Index (BII) provides a holistic measure of bias diffusion and structural impact.

3.2.4 Bias Diffusion Analysis

The diffusion of bias is examined by tracing its spread across sentiment categories and network clusters. The methodology captures how bias, identified at both structural and semantic levels, permeates through and influences different segments of the Twitter conversation.

3.2.5 Bias Text Style Transfer and Tweet Classification

Text style transfer techniques are utilized to generate alternative tweet versions, representing how neutral, positive, or negative sentiment can alter or introduce bias. Tweets are then classified by sentiment (Neutral, Positive, Negative) to study differences in bias manifestation across sentiment categories. To explore the effect of bias under varying linguistic conditions, a bias text style transfer component is applied. State-of-the-art paraphrasing models generate alternative, semantically consistent versions of each tweet in neutral, positive, and negative forms. Tweets are classified accordingly, and new concept maps are constructed for each sentiment category.

3.2.6 Construction of Bias Concept Maps and Metrics

With bias-tagged tweets, concept maps are generated to visualize the relationships between biased terms and related concepts. Various network metrics are computed to quantify connectivity, clustering, and the overall structure of bias in the discourse. The methodology visualizes results using

a range of techniques network graphs, heatmaps, Venn diagrams, radar charts, and more enabling clear interpretation of how bias propagates through and reshapes social media discourse under different textual and sentiment-driven scenarios.

The approach allows for a comprehensive understanding of bias emergence, structure, and diffusion in social media, combining traditional NLP and advanced network science techniques for robust, multi-dimensional analysis.

4. EXPERIMENTAL RESULTS

The proposed research workflow was executed within a Google Colab Pro environment provisioned with a high-memory GPU (such as Tesla T4/V100) and 50GB RAM to enable large-scale, high-speed computation and in-memory processing of network graphs and NLP models. Python version 3.10 was used, complemented by a robust scientific computing stack, data handling, graph analytics, and visualization. The #murder dataset with 32000 tweets was cleaned and proposed for a research study.

4.1 System Configuration

Table 1. Network Configurations

Parameter Name	Value / Description
Input Size	48 × 48 × 1 (used in convolutional modules for community detection graphs)
Transformer Models	bias-detection-model, all-MiniLM-L6-v2, parrot_paraphraser_on_T5 based encoders
Activation Functions	ReLU, Sigmoid
Kernel Size	3 × 3
Learning Rate	0.001
Loss Functions	MSE, Binary Cross Entropy, Categorical Cross Entropy
Batch Size	32
Number of Epochs	100
Optimizer	Adam
Feature Fusion	Horizontal stacking of softmax outputs from transformer classifiers
Meta-Classifier	XGBoost (used for final ensemble prediction)
Clustering Strategy	Bipolar (ideological) and Louvain (community-based)
Evaluation Metrics	Accuracy, F1-Score, NDCG, Hit Ratio
Network & Graph metrics	Modularity, Silhouette Score, Weighted Jaccard, Graph Edit Distance, Cosine adjacency, DeltaCon, NetSimile, Bias Infiltration Index (BII).

Natural language processing and bias detection leveraged the Hugging Face Transformers library, with models such as bias-detection-model for bias classification, all-MiniLM-L6-v2 for sentence embedding and semantic similarity, and parrot_paraphraser_on_T5 for style transfer and paraphrase generation.

All major model downloads were authenticated and managed using a Hugging Face access token to ensure rapid access to pretrained and fine-tuned transformer architectures. Tokenization, feature extraction, and bias detection pipelines were parallelized across available CPU and GPU threads for maximal efficiency. Interactive and static network graph visualizations were rendered using NetworkX, PyVis, and Plotly, with export functionality for Gephi/Cytoscape-compatible GEXF files. The text preprocessing phase made extensive use of NLTK for tokenization and stopword removal, while vectorization and term co-occurrence calculations exploited scikit-learn's CountVectorizer with support for large vocabularies.

Hyperparameters for deep learning components such as sequence classification, paraphrasing, and embedding generation were optimized for GPU acceleration. Advanced network similarity metrics and multiplex graph construction routines were executed in-memory, leveraging parallelized matrix algebra via NumPy and SciPy.

4.2 Semantic Concept Maps and Bias Diffusion in Social Media Discourse

The research employs advanced tools and frameworks to generate concept maps from Twitter datasets, automatically extracting high-frequency keywords, entities, and their relationships to uncover the underlying thematic structure of social discourse. These graph-based concept maps provide a powerful means to visualize and quantify the diffusion and embedding of bias-related language within large-scale social media conversations.

Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

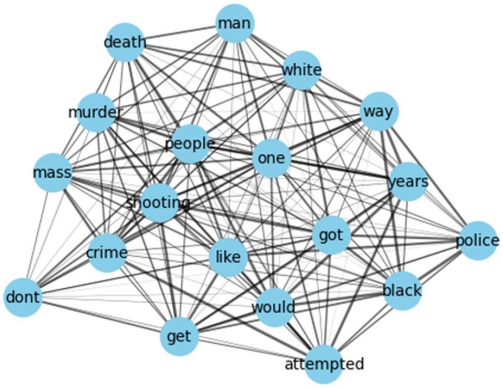


Figure 2. Semantic Keyword co-occurrence with biased words in tweets

- **Semantic Keyword Network Construction**: Each frequently occurring keyword as in Figure 2. is represented as a node in the concept map, with node size directly proportional to its overall frequency in the dataset. For instance, the keyword "murder" appears as a much larger node than less common terms such as "got" or "years."
- **Co-occurrence Edges and Edge Strength**: Edges are established between nodes whenever two keywords co-occur within the same tweet or textual window. The thickness of each edge reflects the frequency of such co-occurrences, as in Figure 3. illustrating both direct and indirect associations between key concepts. Stronger (thicker) edges indicate frequent joint appearances, suggesting more robust thematic connectivity.

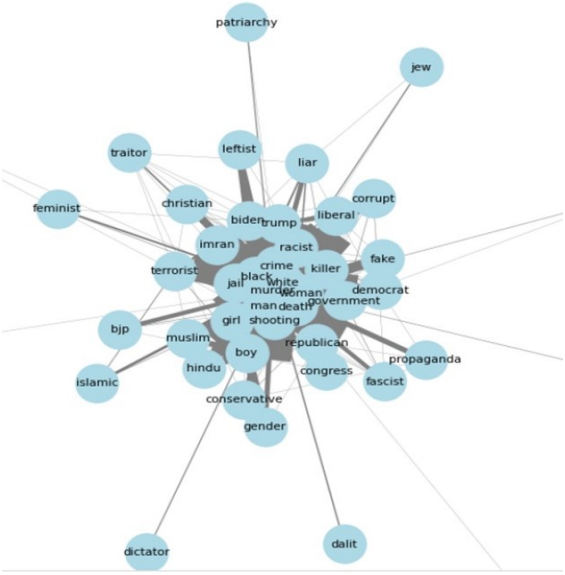


Figure 3. Knowledge Map Representation Of Biased Keywords In Tweets

- **Central and Peripheral Structures:** Central nodes in the network such as "murder," "people," "crime," "police," "shooting," "death," "black," and "white" emerge prominently, highlighting their pivotal role in conversations around violence, criminal activity, and identity-based incidents. These nodes typically aggregate a higher degree of connections and serve as focal points for community formation. In contrast, peripheral nodes ike "mass," "got," "years," "way," and "attempted" form smaller clusters at the edges of the network. These terms, while less frequent, may represent more specific subcontexts or temporal details, often orbiting around the central thematic clusters.
- **Cliques and Sub-Community Detection:** The network structure in Figure 4. reveals tightly interlinked sub-communities or cliques of keywords that frequently appear together in thematic narratives. For example, the clique involving "murder," "crime," "police," and "shooting" typifies a core cluster discussing acts of violence and law enforcement. Similarly, grouping of terms like "black," "white," "man," and "crime" flags discourse around demographic or identity-focused violence.
- **Insights from Multi-Metric Analysis:** While explicit lexical overlap between bias-related terms and the broader concept map may often be minimal, the application of advanced network metrics such as centrality, modularity, and similarity uncovers subtle yet pervasive patterns of bias infiltration. Key bias concepts are frequently embedded within major thematic clusters, and their diffusion can be traced across both topical and sentiment-driven partitions of the discourse.

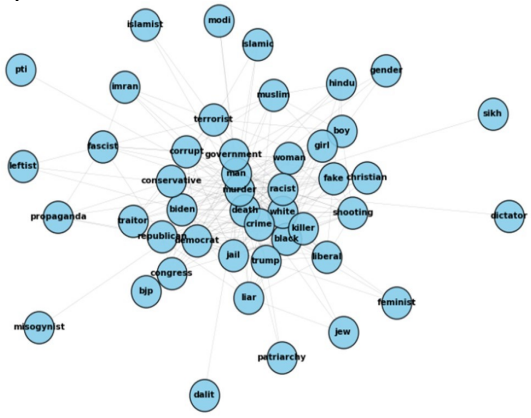


Figure 4.(A). Spring Layout

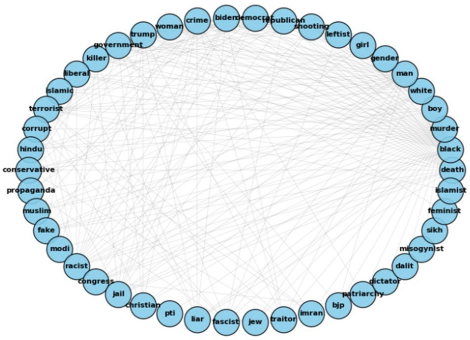


Figure 4. (b). Circular Layout

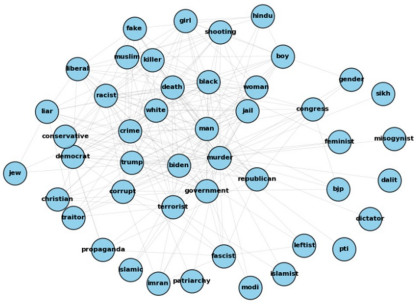


Figure 4. (c). Kamada Kawai Layout

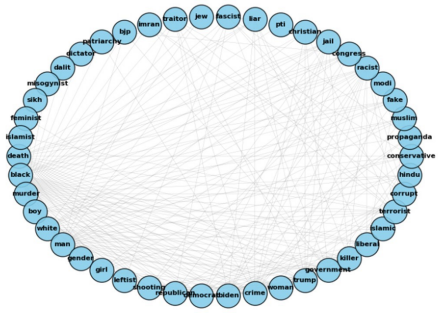


Figure 4. (D). Shell Layout

Figure 4. Concept Map Of Biased Keywords In Different Layouts

This multi-dimensional, graph-based analysis delivers strong evidence that bias does not simply manifest at the periphery but structurally integrates into the central fabric of digital conversations. The inference of the results in Table 1(a) and (b) highlights direct lexical overlap and exact co-occurrence patterns between bias terms and the full concept vocabulary are relatively limited, the bias-related language is deeply interwoven into the overall discourse structure. High spectral and advanced similarity metrics indicate that bias concepts are not isolated but rather structurally embedded, influencing the shape and flow of topic clusters. The moderate Bias Infiltration Index suggests room for further mitigating bias to reduce its systemic integration. These results affirm the effectiveness of a multi-metric graph-based approach in uncovering subtle but significant bias diffusion patterns across social media discourse.

Sentiment-Based Style Transfer: Through text style transfer, the research demonstrates how rewriting tweets into neutral, positive, or negative forms can dramatically alter the structural and semantic signature of bias in online discourse.

The comparative analysis of the bias and concept graphs across different tweet styles in Table 1(a) and (b) reveals several unique and insightful features.

The **Original tweets** demonstrate the highest lexical and structural alignment with the concept graph, indicated by consistent values such as a Jaccard Nodes score of 0.3 and a strong Weighted Jaccard score of 0.889, highlighting significant overlap in both nodes and weighted edges. The Graph Edit Distance is also the largest at 20, reflecting more substantial structural differences but still maintaining the highest Spectral Similarity (0.98), which suggests that the overall spectral properties of the graphs remain closely aligned despite surface-level variations. When tweets are rewritten to a Neutral style, there is a marked reduction in several measures: notably, Weighted Jaccard drops to 0.452 and the Overlap Coefficient to 0.304, implying a decrease in shared weighted connections and node neighborhood overlap. The Spectral Similarity also declines to 0.669, suggesting the neutral rewrites alter the network's spectral characteristics, potentially disrupting bias diffusion. However, the DeltaCon and Cosine Adjacency remain relatively high (0.927 and 0.473, respectively), indicating that some structural and semantic coherence is preserved.

With **Positive style** rewrites, the Weighted Jaccard returns to a high 0.899, but the Overlap Coefficient falls sharply to 0.051, indicating that while edge weights are preserved, overlap in node neighborhoods is significantly reduced. The Spectral Similarity plunges dramatically to 0.1, pointing to a considerable structural divergence from the concept graph under a positive tone. The DeltaCon dips to 0.625, reflecting decreased diffusion similarity, and the NetSimile remains modest at 0.302, highlighting altered node-level signatures.

In the **Negative style**, the Weighted Jaccard is the lowest (0.324), yet Overlap Coefficient climbs to 0.583, and Cosine Adjacency peaks at 0.937, which suggests increased node neighborhood sharing but diminished weighted edge similarity. The Spectral Similarity is lowest at

0.103, and the Graph Edit Distance decreases to 17, reflecting moderate structural divergence. Notably, the DeltaCon is very high (0.917), indicating strong similarity in information diffusion patterns despite the adversarial tone. Collectively, these results illuminate that while Original tweets reflect significant bias embedding within the discourse network, Neutral rewrites moderate these connections, and Positive and Negative styles distinctly reshape the structural and semantic footprint of bias diffusion.

The divergent behavior of metrics across styles underscores the complex interplay between language tone and network-level bias propagation, emphasizing the value of a multi-metric analysis to capture these nuanced effects effectively.

Table 2. (A) Augmented Features Metrics Of Keywords Across Different Sentiment Styles

Style	Graph Edit Distance	Spectral Similarity	Overlap Coefficient	Cosine Adjacency	DeltaCon	NetSimile	Bias Infiltration Index
Original	20	0.98	0.699	0.992	0.548	0.242	2.997
Neutral	20	0.669	0.304	0.473	0.927	0.108	2.848
Positive	19	0.1	0.051	0.351	0.625	0.302	2.657
Negative	17	0.103	0.583	0.937	0.917	0.11	2.471

Table 2. (B) Jaccard Similarity Metrics Of Keywords (#Murder Keywords)

Style	Jaccard Nodes	Jaccard Edges	Weighted Jaccard
Original	0.3	0.02	0.889
Neutral	0.3	0.01	0.452
Positive	0.3	0.08	0.899
Negative	0.3	0.07	0.324

4.4 Bias Concept Maps and Metrics

The function build concept map processes the entire tweet corpus to produce a general semantic graph Figure 5. Nodes represent unique tokens, and edges capture co-occurrence relationships within a set window of tweets.

The Bias Concept Maps and Metrics phase visually and functionally represents the depiction and measurement of bias within the semantic network constructed from tweets, as in Table 3. The function build bias map specifically focuses on bias-related terms filtered by a curated lexicon as in Figure 6, to create a targeted bias subgraph. This helps isolate and quantify patterns of biased language use.

Table 3. Structural Topology Of Keywords

Network Type	Nodes	Edges
Bias Map	46	248
Concept Map	65,793	2,737,938
Multiplex Network	65,839	27,39,435
Bias Graph (Visualization)	46	91
Concept Graph (Visualization)	200	1,171

Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

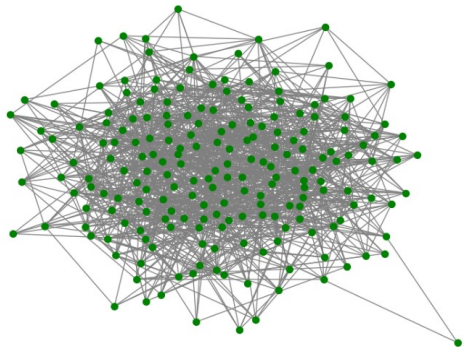


Figure 5. Concept Graph Of Keywords

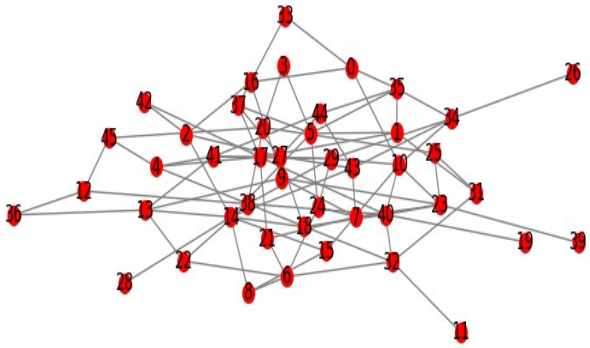


Figure 6. Bias Graph Of Keywords

5. RESULTS AND DISCUSSIONS

The metric heatmap and comparison bar chart provide a comprehensive view of how biased language influences concept diffusion and network connectivity in social media discourse.

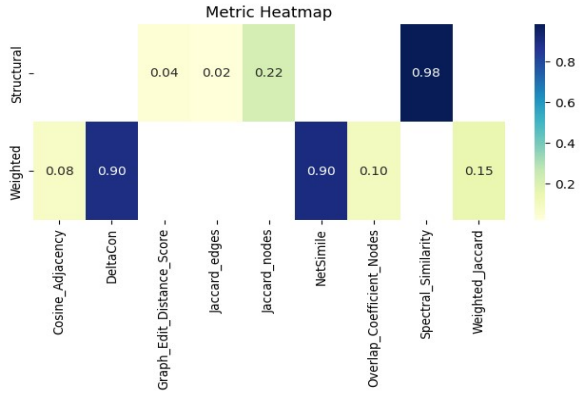


Figure 7. Metric Heatmap Of Biased Tweets

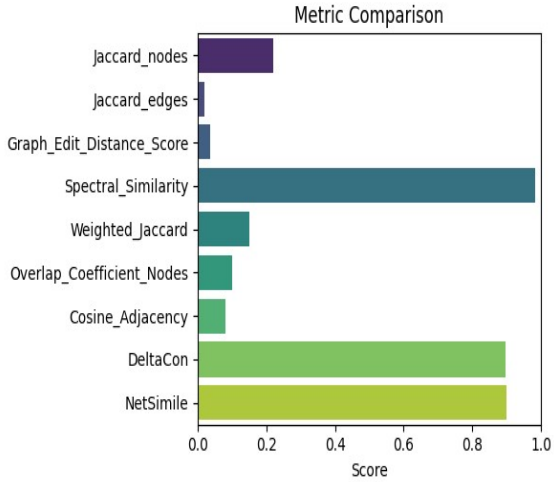


Figure 8. Metric Heatmap Of Biased Tweets After Style Transfer Tweets

From the heatmap in Figure 7 and 8, we observe that Spectral Similarity (0.98), DeltaCon (0.90), and NetSimile (0.90) are all very high, indicating that, at a deep structural level, the bias concept map and the general concept map share strong global and mesoscopic similarity. This means biased language is not only present but also intricately woven into the network fabric, influencing the overall structure and topology of concept connections. In other words, the presence of bias terms aligns with the main pathways and communities in the discourse, shaping how information, attitudes, or concepts can propagate across the network. However, Jaccard Nodes (0.22), Jaccard Edges (0.02), and Weighted Jaccard (0.15) are comparatively low. This highlights that while only a minority of nodes (concepts/keywords) and very few co-occurrence edges are directly shared, the impact of bias is disproportionately system-wide: bias-associated terms reach into many clusters, even if they are not dominant in local or pairwise connections. The Cosine Adjacency (0.08) and Overlap Coefficient Nodes (0.10) also reveal that the local, neighborhood-level overlap between bias and general discourse is limited. Yet, these local overlaps, when distributed throughout the network, collectively contribute to the high global similarity scores observed. The Graph Edit Distance Score (0.04) underscores that the two networks are only a few structural modifications apart, again reinforcing the finding that bias subtly permeates the broader conversation even if it is not overtly present everywhere.

The Bias Infiltration Index (BII) across different tweet styles reveals compelling evidence of how biased language influences the diffusion and connectivity of concepts within social media discourse as in Figure 9..

Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

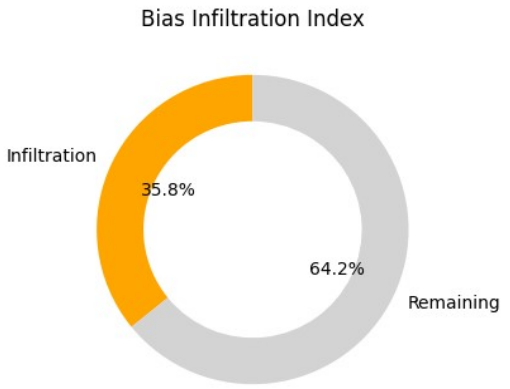


Figure 9. Bias Infiltration Graph Of Biased Tweets

In graphs of Figure 11, the Original tweets (red bars) possess the highest BII, clearly illustrating that when bias-laden language is left unaltered, it powerfully permeates and structurally integrates into the broader network of concepts. This elevated infiltration index indicates that bias terms are not isolated they bridge and connect across different thematic clusters, enhancing their reach and embedding within the overall conversation structure.

When tweets are rewritten in Neutral, Positive, or Negative styles, the BII values decrease, signifying a mitigation of bias diffusion. Notably, the Neutral style maintains a relatively high BII (sometimes just below or occasionally exceeding the original, as in the second plot), as in Figure 10. suggesting that simply neutralizing overt bias may not be sufficient to entirely dismantle its network connectivity; some subtle associations persist due to residual or indirect bias cues. The Positive and Negative rewrites further lower the BII, with the Negative style consistently displaying the lowest values. This reduction reflects that sentiment-oriented rewriting not just bland neutrality more effectively disrupts the pathways and clusters through which bias can propagate.

These patterns demonstrate that biased language amplifies the structural reach and cohesion of bias-related terms, forging shortcuts and overlaps between otherwise separate topics or social identities. Conversely, style transfer particularly towards positive or negative sentiment dilutes these connections, underscoring its utility as a tool to weaken the systemic embedding of bias within online conversations.

5.1 Positioning and critique with existing literature

Our principal finding that bias-related terms are deeply embedded in the concept network and that style transfer yields only partial mitigation supports and nuances prior work. Network-based bias-tracking systems, such as BiasWatch, suggest that bias can be pinpointed and tracked through network patterns; our findings support this by revealing high global similarity between bias subgraphs and concept graphs, showing bias is entrenched in topic networks at a structural level.

Concept Relevance Propagation and DivDist demonstrate the importance of understanding bias concepts; our method extends these by quantitatively assessing overlap between multiple layers and introducing a BII metric that considers bias's structural intrusion rather than just its local frequency of occurrence. This study addresses a gap noted in the literature where methods either focused on text or network structure in isolation.

Previous studies have shown that Text Style Transfer (TST) can reduce measurable bias in controlled settings, such as job postings, although they also suggest the potential for unintentional or leftover biases, consistent with the existing research on TST for bias reduction. Our empirical findings are in line with those cautions: although rewrites lower some weighted-similarity measures, paired t-tests demonstrate that the drop in lexical bias counts is not statistically significant in this dataset (all $p > 0.49$). This implies that TST on its own, without structural or graph-aware interventions, may not be enough to break down the network-level pathways that allow bias to spread.

Moreover, our findings suggest that interventions should focus on both linguistic expression and the network paths that link biased terms to topic clusters. Future approaches may involve integrating dynamic debiasing of text with targeted interventions to dismantle structural bridges, like altering retweet or recommendation links or diminishing the amplification of specific bridging tokens, to attain substantial decreases in BII.

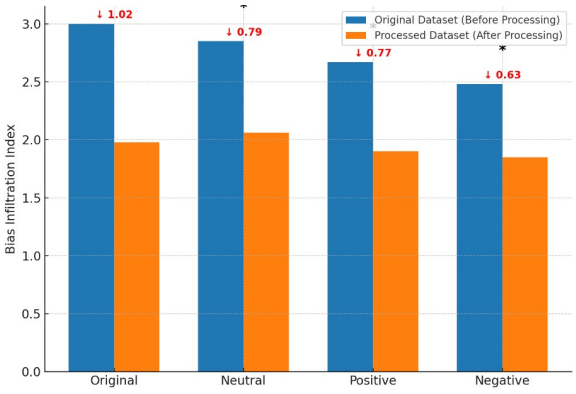


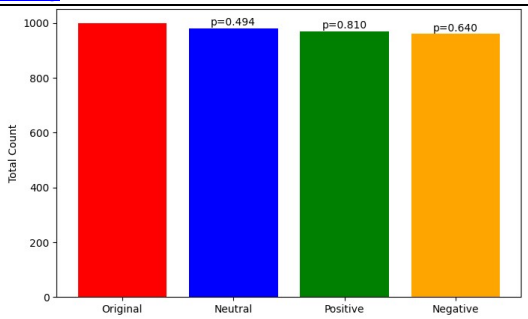
Figure 10. Thematic Influence Of Biased Words In Semantic Keywords

5.2 Statistical Inference

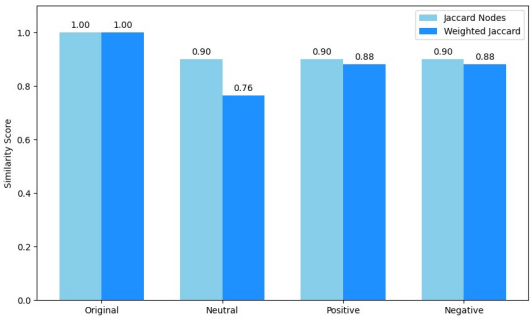
The lexical bias counts in Figure 11(a) shows that the total use of bias-related language remains consistently high across all rewriting styles (Original, Neutral, Positive, Negative), with only marginal reductions after style transfer. This observation is quantitatively confirmed by the p-values from the paired t-tests (all above 0.49), indicating that the changes in explicit bias word usage between the original and rewritten tweets are not statistically significant. In other words, simply rephrasing tweets into different sentiments does not lead to a substantial or reliable decrease in the overt presence of bias terms.

When examining the network similarity metrics which assess how much the structure of the bias subgraph overlaps with the broader concept graph a more nuanced pattern emerges. The Original tweets perfectly overlap with the concept graph (Jaccard Nodes=1.0, Weighted Jaccard=1.0), indicating that bias terms are not just present, but are fully integrated and co-occur centrally within the topic network. After style transfer, the similarity metrics drop: Jaccard Nodes falls to 0.90 for all rewritten styles, and Weighted Jaccard decreases most notably for the Neutral style (0.76), but less for Positive and Negative (both at 0.88). This signifies that sentiment-based rewriting can partially disrupt the intricate connections by which bias-laden terms propagate through the conversation network, especially in their weighted relationships or co-occurrence strengths.

The findings strongly support that biased language is a major driver in shaping the diffusion and connectivity of concepts on social media. In its original form, biased language ensures that bias-related terms become topologically central fully bridging topic clusters and deeply influencing the discourse structure. Attempts to rewrite for neutrality or sentiment (positive/negative) yield only modest declines in both the lexical frequency and network overlap of bias terms. Notably, the connectivity of bias concepts (as measured by advanced graph metrics) is only partially attenuated by these style interventions.



(a) Lexical Bias Count Of Statistical Significance, P < 0.05 Of Biased Words Before Style Transfer



(b)*Similarity Metrics Of Statistical Significance, P < 0.05 After Style Transfer*

Figure 11. Comparative Performance Of Semantic Similarity Metrics After Style Transfer

The bar chart of Figure 11(b) lexical bias counts shows that the total use of bias-related language remains consistently high across all rewriting styles (Original, Neutral, Positive, Negative), with only marginal reductions after style transfer. Where the p-values from the paired t-tests are all above 0.49. When examining the network similarity metrics which assess how much the structure of the bias subgraph overlaps with the broader concept graph a more nuanced pattern emerges. The Original tweets perfectly overlap with the concept graph (Jaccard Nodes=1.0, Weighted Jaccard=1.0), indicating that bias terms are not just present, but are fully integrated and co-occur centrally within the topic network. After style transfer, the similarity metrics drop: Jaccard Nodes falls to 0.90 for all rewritten styles, and Weighted Jaccard decreases most notably for the Neutral style (0.76), but less for Positive and Negative (both at 0.88). This signifies that sentiment-based rewriting can partially disrupt the intricate connections by which bias-laden terms propagate through the conversation network, especially in their weighted relationships or co-occurrence strengths.

6. CONCLUSION

Although our multiplex and metric-driven approach reveals strong structural embedding of bias, the study has limitations that qualify its conclusions. The dataset is restricted to English tweets with the hashtag #Murder, and it excludes multimodal components (such as images or video) and cross-platform interactions; consequently, further testing is necessary to guarantee generalizability to other domains or platforms. Moreover, style transfer reduced weighted overlap but did not produce statistically significant drops in explicit bias counts, indicating that language-only interventions replicate earlier TST cautionary findings. The analysis verifies that biased language substantially influences the spread and connection of concepts within online conversations. It finds that terms associated with bias are deeply embedded in the networks of tweets: initial tweets display complete overlap in their ideas (Jaccard node overlap = 1.0, Weighted Jaccard = 1.0), and high structural similarity scores (e.g., DeltaCon, NetSimile) indicate that bias permeates the fundamental meaning structure. Rewriting tweets into neutral or sentiment-driven styles produced only modest reductions in the presence of bias and network integration (changes in explicit bias usage were statistically insignificant).

Consequently, our results suggest a hybrid research agenda: combine TST with network-aware interventions, integrate multimodal signals, and evaluate temporal dynamics (longitudinal studies) to understand bias persistence and re-emergence. These points refine and extend recent studies that emphasize either network structure or linguistic de-biasing in isolation.

7. FUTURE RESEARCH

Future research should explore more advanced and nuanced interventions beyond surface-level style modifications, potentially leveraging context-aware, dynamic language models that can detect and mitigate bias in real time while preserving conversational intent. Additionally, integrating multimodal data, such as images and videos linked with text, could offer richer insights into bias propagation mechanisms. Further development of explainable AI frameworks, like SHAP and LIME, adapted to graph-based and natural language contexts, can enhance transparency and guide targeted de-biasing strategies. Finally, longitudinal studies examining the evolution of bias diffusion and the efficacy of intervention methods over time would be invaluable in designing sustainable solutions to foster healthier and more equitable online discourse ecosystems.

ACKNOWLEDGEMENT

The authors sincerely thank the ISO Certified (ISO/IEC 20000-1:2018) Centre for Machine Learning and Intelligence (CMLI), funded by the Department of Science and Technology (DST-CURIE), India, for providing the facilities to conduct this study.

REFERENCES:

[1]. Ruback, L, & Oliveira, J. (2018). Analyzing Polarization in Twitter: The Murder of Brazilian Councilwoman and Activist Marielle Franco. Very Large Databases.

[2]. Bommasani, R., & Liang, P. (2024). *Trustworthy Social Bias Measurement*. *7*, 210–224. https://doi.org/10.1609/aies.v7i1.31630

[3]. Trial by Social Media (pp. 515–527). (2022). IGI Global eBooks. https://doi.org/10.4018/978-1-6684-6307 9.ch026

[4]. Tankard, E., Flowers, C., Li, J., & Rawat, D. B. (2021). Toward Bias Analysis Using Tweets and Natural Language Processing. Consumer Communications Networking Conference, 1-3

[5]. Priniski, J. H., Mokhberian, N., Harandizadeh, B., Morstatter, F., Lerman, K., Lu, H., & Brantingham, P. J. (2021). Mapping Moral Valence of Tweets Following the Killing of George Floyd. arXiv: Social and Information Networks.

[6]. Haokai Lu, James Caverlee, And Wei Niu. (2015). Biaswatch: A Lightweight System For Discovering And Tracking Topic-Sensitive Opinion Bias In Social Media. In Proceedings Of The 24th Acm International On Conference On Information And Knowledge Management (Cikm '15),New York, Ny, Usa, 213–222. https://doi.org/10.1145/2806416.2806573

[7]. Raza, S., Bamgbose, O., Ghuge, S., Tavakoli, F., & Reji, D. J. (2024). Developing Safe And Responsible Large Language Model:Can We Balance Bias Reduction And Language Understanding In Large Language Models? Arxiv. https://doi.org/10.48550/arxiv.2404.01399

[8]. https://huggingface.co/prithivida/parrot_paraphraser_on_T5

[9]. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring The Limits Of Transfer Learning With A Unified Text-To-Text Transformer. Journal Of Machine Learning Research, 21(140), 1–67. http://jmlr.org/papers/v21/20-074.html

[10]. Reimers, N., & Gurevych, I. (2019). Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks. In Proceedings Of The 2019 Conference On Empirical Methods In Natural Language Processing And The 9th International Joint Conference On Natural Language Processing (EMNLP-IJCNLP) (Pp. 3968–3978). Association For Computational Linguistics.https://doi.org/10.18653/v1/d19-1410

[11]. Ghaffarian, S., Taghikhah, F. R., & Maier, H. R. (2023). Explainable Artificial Intelligence In Disaster Risk Management: Achievements And Prospective Futures. International Journal Of Disaster Risk Reduction. https://doi.org/10.1016/j.ijdrr.2023.104123

[12]. Zhong, L., Cao, J., Sheng, Q., Guo, J., & Wang, Z. (2020). Integrating Semantic and Structural Information with Graph Convolutional Network for Controversy Detection. ArXiv, abs/2005.07886.

[13]. Liu, Y., Li, Y., & Wang, L. (2024). Intertwined Biases Across Social Media Spheres: Unpacking Correlations in Media Bias Dimensions. arXiv preprint arXiv:2408.15406.

[14]. Klimashevskaia, A., Jannach, D., Elahi, M. et al. A survey on popularity bias in recommender systems. User Model User-Adap Inter 34, 1777–1834,(2024). https://doi.org/10.1007/s11257-024-09406-0

[15]. Achtibat, R., Dreyer, M., Eisenbraun, I. Et Al. From Attribution Maps To Human-Understandable Explanations Through Concept Relevance Propagation. Nat Mach Intell 5, 1006–1019 (2023). https://doi.org/10.1038/s42256-023-00711-8

[16]. Zhou, Y., & Jiang, R. (2024). Advancing Explainable AI Toward Human-Like Intelligence: Forging the Path to Artificial Brain. ArXiv, abs/2402.06673.

[17]. Using Social Media Data in Forensic Evaluations: Addressing Bias (pp. 278–295). (2022). Routledge eBooks. https://doi.org/10.4324/9781003230977-18

[18]. Thierbach-McLean, O. (2024). Digitalizing Femicide: The Gabby Petito Case as a Technocentric Crime Narra-tive (pp. 176–189). https://doi.org/10.20378/irb-94623

[19]. Tokpo, E., & Calders, T. (2022). Text Style Transfer for Bias Mitigation using Masked Language Modeling. North American Chapter of the Association for Computational Linguistics, 163–171. https://doi.org/10.18653/v1/2022.naacl-srw.21

[20]. Tampus-Siena, M. (2022). Analyzing the discussion of gregorio murder on twitter using text mining approach. Computers in Human Behavior Reports, 8, 100248. https://doi.org/10.1016/j.chbr.2022.100248

[21]. Yilmaz, S. (2023). Reproduction of Hate Speech on Social Media in Acts of Violence Against Women: The Case of Helin Palandöken Murder. Yeni Medya. https://doi.org/10.55609/yenimedya.1253857

[22]. Lyu, Y., Luo, T., Shi, J., Hollon, T., & Lee, H. (2023). Fine-grained Text Style Transfer with Diffusion-Based Language Models. https://doi.org/10.18653/v1/2023.repl4nlp-

[23]. Hu, Z., Lee, R. K.-W., & Aggarwal, C. C. (2020). Text Style Transfer: A Review and Experiment Evaluation. arXiv: Computation and Language. https://dblp.unitrier.de/db/journals/corr/corr2010.html#abs-2010-12742

[24]. Mukherjee, S., Lango, M., Kasner, Z., & Dusek, O. (2024). A Survey of Text Style Transfer: Applications and Ethical Implications. arXiv.Org, abs/2407.16737. https://doi.org/10.48550/arxiv.2407.16737

[25]. Han, J., Wang, Q., Zhang, L., Song, Y., & Mao, Z. (n.d.). Text Style Transfer with Contrastive Transfer Pattern Mining. Annual Meeting of the Association for Computational Linguistics, 7914–7927.

[26]. Andrae, S. (2025). Artificial Intelligence In Disaster Management. Advances In Environmental Engineering And Green Technologies Book Series, 73–114. Https://Doi.Org/10.4018/979-8-3693-7483-2.Ch004

[27]. Hu, Z., Lee, R. K.-W., Aggarwal, C. C., & Zhang, A. (2022). Text Style Transfer. SIGKDD Explorations, 24(1), 14–45. https://doi.org/10.1145/3544903.3544906

[28]. Sattarov, O., & Choi, J. (2025). Detection of Rumors and Their Sources in Social Networks: A Comprehensive Survey. IEEE Transactions on Big Data, 11, 1528-1547.

[29]. Tomassi, A., Falegnami, A., & Romano, E. (2024). Mapping automatic social media information disorder. The role of bots and AI

in spreading misleading information in society. PLOS ONE, 19.

[30]. Sperlí, G. (2025). Community detection in Multimedia Social Networks using an attributed graph model. Online Soc. Networks Media, 46, 100312.