International Journal of
Computational Linguistics Research

DLINE JOURNALS

Print ISSN: 0976-416X
Online ISSN: 0976-4178

IJCLR 2026: 17 (2)
https://doi.org/10.6025/ijclr/2026/17/2/78-99

Modeling and Analyzing Engagement Dynamics of Misleading and Authentic Content on Reddit Using Linguistic and Machine Learning Approaches

Hathairat Ketmaneechairat
Faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
hathairat.k@cit.kmutnb.ac.th

ABSTRACT

The rapid proliferation of misleading information on social media poses significant challenges to digital ecosystems, driven by sensational narratives, emotional framing, and strategic engagement tactics. Existing research often examines linguistic patterns, user engagement, and propagation dynamics in isolation, limiting comprehensive understanding. This study introduces a unified analytical framework that integrates linguistic, behavioral, and contextual features to model and classify misleading versus authentic content on Reddit. Analyzing a dataset of 2,344 posts across twelve subreddits, we engineered 86 multidimensional features encompassing linguistic structure, sentiment, stylistic markers, clickbait indicators, and TF-IDF representations. A Linear Support Vector Machine (LinearSVC) was employed for multi-class classification across true, satire, imposter, and misleading categories. Descriptive and statistical analyses revealed that authentic content attracts the highest average engagement, while misleading posts exhibit significantly greater verbosity and clickbait prevalence, strategically mimicking credible narratives. The baseline classifier achieved 88.7% weighted accuracy, effectively distinguishing satire and imposter content, yet struggled with misleading posts due to their deliberate lexical overlap with authentic news. Inferential tests confirmed statistically significant differences in engagement metrics, content length, and clickbait usage across categories. These findings demonstrate that virality is decoupled from factual accuracy and highlight the limitations of traditional machine learning in detecting nuanced deception. Future research should incorporate transformer-based architectures, temporal engagement modeling, and network propagation analysis to enhance robust misinformation detection in dynamic online communities.

Keywords: Misinformation Detection, Reddit Engagement, Linguistic Features, Machine Learning Classification, Content Authenticity, Clickbait Indicators, Feature Engineering, Social Media Dynamics

Received: 21 August 2025, Revised 31 October, Accepted 19 November 2025

1. Introduction

The proliferation of misleading information on social media platforms has emerged as a critical challenge in the digital information ecosystem. The success of fake news propagation is often attributed to intentionally exaggerated narratives, emotionally charged language, compelling imagery, and strategically designed clickbait mechanisms [1] . These elements are frequently combined to maximize user engagement and virality. In addition to content characteristics, psychological factors significantly influence the spread of misinformation, as users may engage with such content, deliberately or unintentionally, due to cognitive biases and social influence [2, 3] .

Interestingly, while misleading posts may initially receive fewer responses compared to authentic content, they tend to accumulate higher engagement over time by leveraging sensational elements such as anger, conflict, and moral framing [4]. This phenomenon is further amplified by the existence of organized fake engagement services, which operate as structured economic ecosystems involving multiple actors across a supply chain [5, 6] .

2. Review of Earlier Studies

2.1 Role of Multimedia and Context in Misinformation

Visual content plays a crucial role in shaping user perception and engagement. Fake images may either be digitally manipulated or authentic images repurposed in misleading contexts. Such misuse includes presenting outdated images as current events or interpreting visuals with incorrect narratives [7,8]. This highlights the importance of contextual integrity in evaluating multimedia content.

2.2 Temporal Dynamics of Engagement

Recent studies emphasize the importance of temporal analysis in understanding the spread of misinformation. Temporal behavioral patterns provide insights into how deceptive content evolves and sustains engagement over time [9]. The dynamics of user interaction with content are influenced by various factors, including content attributes and platform-specific mechanisms.

Kuo [10] explored the relationship between post attributes and user engagement in the context of correcting health misinformation, demonstrating that engagement is driven not only by content accuracy but also by presentation and framing. Similarly, Candogan [11] examined the trade-offs between engagement and misinformation under different dissemination mechanisms, identifying optimal conditions for balancing reach and reliability.

From a media studies perspective, Castaldo [12] introduced the concept of “over-accelerated attention regimes,” where a small subset of content captures disproportionate attention but fails to sustain it, leading to fragmented and short-lived public discourse (Maria Castaldo). This framework provides a valuable lens for understanding the ephemeral nature of viral misinformation.

2.3 Misinformation Correction and User Response

The effectiveness of misinformation correction strategies has been widely studied across different contexts. Prior research has investigated how variables such as message complexity (simple vs. detailed explanations),

source credibility, and content characteristics (quality, intensity, emphasis) influence user acceptance of corrections [13, 19]. These findings suggest that both the structure and delivery of corrective information play a critical role in mitigating misinformation.

2.4 Approaches to Fake News Detection

Fake news detection methodologies can be broadly categorized into content-based and social context-based approaches. Content-based methods primarily focus on textual features, including linguistic patterns and semantic structures. However, these approaches may face limitations in social media environments due to the informal and dynamic nature of user-generated content [20].

In contrast, social context-based approaches leverage user behavior, interaction patterns, and network structures to identify misinformation [1]. Several advanced models have been proposed in this domain. For instance, Yang et al. [21] introduced an unsupervised detection framework based on hierarchical user engagement data. Liu and Wu [22] developed a propagation-based model using recurrent and convolutional neural networks to capture the diffusion patterns of fake news.

Dordevic et al. [23] proposed a comprehensive framework incorporating 27 variables across users, content, and network structures, including graph based features such as edges, vertices, and infection dynamics. Multimodal approaches have also gained prominence, with Jin et al. proposing an attention-based recurrent neural network (att-RNN) that integrates textual and visual features using LSTM and pre-trained VGG19 representations [24]. Similarly, Wang et al. [25] developed the Event Adversarial Neural Network (EANN), which extracts event invariant features to improve generalization across different misinformation scenarios [25]. Additionally, deep learning techniques have been applied to detect and generate deepfakes, further expanding the scope of misinformation research [26].

2.5 Machine Learning Techniques for Detection

A wide range of machine learning algorithms has been employed to detect fake news and related phenomena. Feature-based detection methods have been extensively used to identify anomalous patterns in user behavior and content characteristics [27, 28, 29]. Traditional classifiers such as Support Vector Machines (SVM) have also demonstrated effectiveness in misinformation detection tasks [30, 31].

Advanced approaches include sequential mining optimization techniques [32] and hybrid frameworks that combine machine learning with evolutionary algorithms to address dataset bias and improve robustness. For example, Faith et al. applied machine learning methods to detect fake accounts on Instagram and proposed a genetic algorithm based approach to mitigate bias [30].

2.6 Reddit as a Platform for Engagement Analysis

Reddit provides a unique environment for studying engagement dynamics due to its community driven structure and voting mechanisms. Users can upvote or downvote content, influencing its visibility and reach. These technical features significantly shape communication practices and content dissemination [33, 34, 35]

Each Reddit community, or subreddit, operates under a defined set of rules and norms that govern user interactions [36]. These norms vary across communities but also adhere to certain platform wide conventions [37, 38] . Compared to other social media platforms, Reddit has a smaller but highly engaged user base, with

approximately 15% of Americans using the platform [39].

Empirical studies have highlighted the heterogeneity of engagement patterns across different subreddits. Horne [40] analyzed a large dataset of comments from multiple subreddits, incorporating sentiment, relevance, and content-based features tailored to Reddit's structure. Similarly, Yu [41] (2024) demonstrated that various features have differing impacts on discussion structures, emphasizing the complexity of interaction dynamics.

The study of engagement dynamics in misleading and authentic content requires an interdisciplinary approach that integrates linguistic analysis, machine learning, and social context modeling. The interplay between content characteristics, user psychology, temporal patterns, and platform-specific features plays a crucial role in shaping the spread of misinformation. By leveraging advanced computational techniques and understanding platform dynamics such as those on Reddit, researchers can develop more effective strategies for detecting, analyzing, and mitigating the impact of fake news.

Despite extensive research on fake news detection, limited work has systematically integrated engagement dynamics, linguistic features, and subreddit-level behavior within a unified analytical framework. Furthermore, the interplay between content structure and user engagement in distinguishing misleading and authentic content remains underexplored, particularly in community-driven platforms such as Reddit.

This study makes the following contributions:

- Proposes a unified framework integrating linguistic, engagement, and contextual features
- Provides a comparative analysis across four content types
- Develops a predictive model for misinformation classification
- Identifies structural and behavioral patterns distinguishing misleading content

Existing research highlights three major dimensions of misinformation analysis: (i) content-based linguistic modeling, (ii) social-context and propagation-based approaches, and (iii) engagement and temporal dynamics. However, these dimensions are often studied in isolation. This fragmentation limits the ability to comprehensively understand how misleading content simultaneously leverages linguistic strategies and platform specific engagement mechanisms. To address this gap, the present study integrates these perspectives into a unified analytical framework.

3. Methodology and Experimental Setup

3.1 Research Framework Overview

This study employs a hybrid analytical framework that integrates linguistic analysis, engagement modeling, and machine learning–based classification to investigate the dynamics of misleading and authentic content on Reddit. The methodology is designed to capture both descriptive behavioral patterns and predictive distinctions across content categories. The overall workflow is structured into six sequential stages: data acquisition, preprocessing, feature engineering, analytical modeling, classification, and evaluation. These stages are systematically organized within a modular pipeline to ensure reproducibility and scalability. The complete system architecture of the proposed framework is illustrated in Fig. 1.

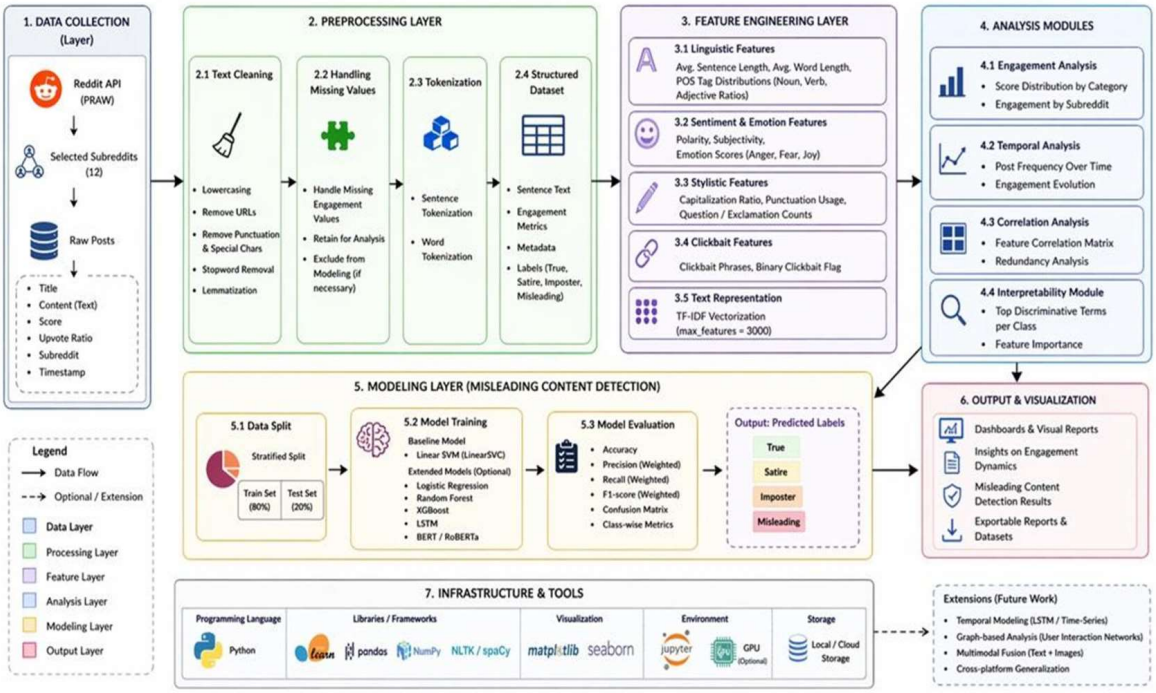

Figure 1. System Architecture for Modeling and Analyzing Engagement Dynamics of Misleading and Authentic Content on Reddit

3.2 System Architecture

The system architecture comprises multiple interconnected layers, each responsible for a distinct stage of data transformation and analysis. The pipeline begins with raw data extraction from Reddit and progresses through structured processing, feature construction, analytical modules, and predictive modeling. As depicted in Fig. 1, the architecture is organized into six primary layers. The Data Collection Layer acquires raw Reddit data via API-based extraction from 12 subreddits spanning diverse content categories, capturing post titles and content, engagement metrics such as scores and upvote ratios, and metadata, including subreddit identifiers and timestamps. The Preprocessing Layer transforms raw textual data into a structured format through text normalization, URL and noise removal, stopword filtering, lemmatization, and tokenization at both sentence and word levels. Missing engagement values are retained for descriptive analysis but excluded from predictive modeling where necessary. The Feature Engineering Layer constructs a comprehensive feature space encompassing linguistic attributes, sentiment polarity and subjectivity, emotion scores, stylistic markers, clickbait indicators, and TF-IDF text representations, with a maximum of 3,000 features. The Analysis Modules layer facilitates exploratory and statistical investigations, including engagement distribution across categories, subreddit-level comparisons, temporal activity trends, feature correlation analysis, and interpretability assessments through discriminative terms. The Modeling Layer performs multi-class classification into true, satire, imposter, and misleading content categories, utilizing a Linear Support Vector Machine trained on TF-IDF features as the baseline, while maintaining extensibility for advanced models such as Random Forest, XGBoost, LSTM, and transformer-based architectures. Finally, the Output and Visualization Layer generates classification results, confusion matrices, feature importance visualizations, and analytical dashboards.

3.3 Dataset Description

The dataset comprises two thousand three hundred forty-four Reddit posts collected from twelve distinct subreddits, categorized into four classes: satire, true, imposter, and misleading content. This classification

scheme facilitates a comparative investigation of the dynamics of authentic and deceptive information. The dataset incorporates both content-level features and engagement-level metrics, thereby supporting a multidimensional analytical approach. Notably, missing engagement values are present in certain subreddits, particularly within AI-generated communities, and are carefully accounted for during data interpretation.

3.4 Data Preprocessing

Data preprocessing is applied to ensure the consistency and quality of input features. The procedure involves the removal of noise such as URLs and special characters, conversion of text to lowercase, tokenization followed by lemmatization, and the elimination of stopwords. Missing engagement values are retained for statistical analysis but systematically excluded during supervised model training to prevent bias.

3.5 Feature Engineering

A total of 86 features are generated to comprehensively represent the dataset's textual and stylistic characteristics. Linguistic features capture structural properties, including average sentence length, average word length, and part-of-speech tag distributions. Sentiment and emotion features quantify the emotional tone through polarity, subjectivity, and scores for anger, fear, and joy. Stylistic features reflect writing conventions by measuring capitalization ratios, punctuation frequency, and the presence of question or exclamation markers. Clickbait indicators are derived as binary and frequency-based metrics to identify attention-driven patterns. For text representation, TF-IDF vectorisation is employed to convert textual data into numerical form, utilising up to 3,000 features to produce a sparse representation suitable for high-dimensional modelling.

3.6 Experimental Design

The experimental design formulates the classification task as a multi-class problem in which each post is assigned to one of four predefined categories. In the classification, each post belongs to one of four categories:

$$Y \in \{\text{True, Satire, Imposter, Misleading}\}$$

$$f: \boldsymbol{X} \rightarrow Y, Y \in \{True, Satire, Imposter, Misleading\}$$

To ensure robust evaluation, the dataset is partitioned into a training set comprising 80% of the samples and a testing set comprising the remaining 20%. Stratified sampling is applied during the splitting process to preserve the original class distribution across both subsets.

3.7 Model Development

The primary model utilized in this study is a Linear Support Vector Machine, selected for its established effectiveness in text classification, computational efficiency in high-dimensional spaces, and robustness to sparse feature representations. To enhance model robustness and generalizability, the experimental testbed also supports implementing alternative algorithms, including Logistic Regression, Random Forest, XGBoost, LSTM-based sequence models, and transformer architectures such as BERT. LinearSVC is selected due to its effectiveness in high-dimensional sparse text representations and its strong generalization performance in text classification tasks compared to probabilistic models.

3.8 Evaluation Metrics

Model performance is assessed using standard classification metrics, namely accuracy, weighted precision, weighted recall, and weighted F1-score. Additionally, a confusion matrix is generated to facilitate detailed

error analysis, and class-wise performance metrics are reported to provide granular insights into model behavior across each content category.

3.9 Experimental Workflow

The complete experimental pipeline follows a structured and sequential workflow, beginning with data collection and progressing through preprocessing, feature engineering, descriptive analysis, model training, performance evaluation, and final interpretation of results.

3.10 Implementation Details

The system is implemented in Python, leveraging a suite of established libraries including scikit-learn for machine learning, pandas and NumPy for data manipulation, NLTK or SpaCy for natural language processing, and Matplotlib for visualization. Development and experimentation are conducted within a Jupyter Notebook environment, with optional GPU acceleration available for deep learning models. Reproducibility is ensured through the use of fixed random seeds and a standardized preprocessing pipeline.

The proposed methodology integrates content analysis, engagement modeling, and machine learning classification within a unified framework. By combining linguistic, emotional, and behavioral features, the system achieves both interpretability and predictive accuracy, establishing a robust foundation for analyzing misinformation dynamics on Reddit.

The computational complexity of the model is primarily governed by TF-IDF vectorization and linear classification, making the framework scalable to larger datasets.

4. Data Analysis and Results

4.1 Dataset Overview and Analytical Context

The dataset comprises 2,344 Reddit posts collected from 12 subreddits, categorized into four content types: *satire*, *true*, *imposter content*, and *misleading content*. This categorization enables a comparative analysis of diverse information types, ranging from authentic news to AI-generated and deceptive narratives.

The dataset integrates both engagement metrics (e.g., score, upvote ratio) and content-based features (e.g., word count, linguistic attributes), facilitating a multidimensional investigation of content performance and structure. However, missing values are observed in key variables, including score (362 instances) and upvote_ratio (763 instances), particularly within AI-generated subreddits. These missing entries are considered during interpretation and represent an inherent limitation of the dataset.

4.2 Descriptive Analysis

4.2.1 Category-Level Patterns

Table 1 presents aggregated statistics across the four content categories, capturing variations in engagement, content length, and stylistic characteristics.

The analysis reveals several important patterns. The *True* category exhibits an exceptionally high average score (~37,033), substantially exceeding other categories, indicating that authentic news content attracts significantly higher engagement. However, the large discrepancy between the mean and median suggests

strong positive skewness, implying that a small number of viral posts disproportionately influence the average.

Category	Posts	Unique Subreddits	Avg Score	Median Score	Avg Upvote Ratio	Avg Title Words	Avg Content Words	Avg Clickbait Markers
Satire	753	3	241.288	12	0.976	12.094	269.174	0.316
True	661	4	37033.113	40334	0.925	13.427	710.198	0.463
Imposter Content	538	4	1404.881	1200.5	0.987	3.429	222.270	0.294
Misleading Content	392	1	6286.036	5182.5	0.966	11.526	755.571	0.709

Table 1. Category-wise Summary Statistics

In contrast, *misleading content* demonstrates the highest average content length (~756 words), even surpassing true content. This suggests that deceptive content may rely on verbosity to simulate depth and credibility. Additionally, misleading posts exhibit the highest prevalence of clickbait markers (0.709), reinforcing the association between misleading narratives and attention-driven strategies.

Despite differences in content type, upvote ratios remain consistently high (>0.92) across categories, indicating generally positive community reception regardless of content authenticity.

To further examine the distribution of engagement, Figure 1 presents the spread of post scores across categories.

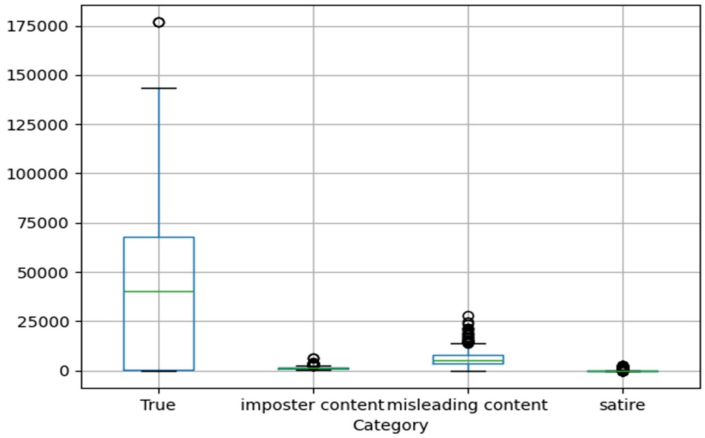

Figure 1. Distribution of post scores across content categories

To further illustrate engagement disparities across content categories, Figure 1 presents the distribution of post scores. The visualization highlights the extreme skewness in the True category, where a small number of highly viral posts drive the average upward, in contrast to the relatively compact distributions observed in satire and imposter content.

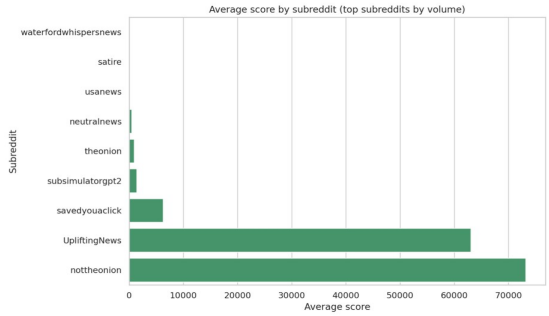

Figure 2. Average Scores by Subreddit

The figure highlights that engagement is highly concentrated within a few dominant subreddits, indicating platform-level inequality in content visibility and reach.

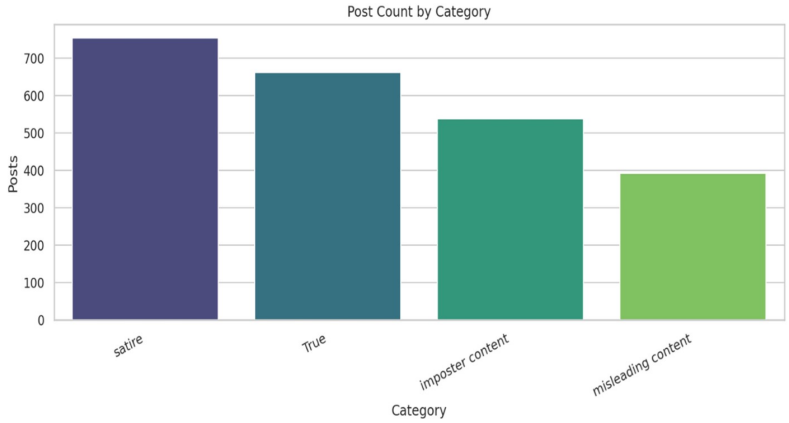

Figure 3. Post count by category

The distribution shows that *satire* and *true* content dominate the dataset, while *misleading content* is relatively underrepresented. This imbalance may influence downstream modeling performance.

The analysis reveals several key patterns.

• Engagement Disparity: The *True* category exhibits overwhelmingly higher average scores (~37,033), suggesting that authentic news attracts significantly greater user engagement compared to other content types.

• Content Length Variation: *Misleading content* has the highest average content length (~755 words), even exceeding *True* content, indicating that deceptive posts may rely on verbosity to enhance credibility.

• Clickbait Prevalence: The highest clickbait marker presence is observed in *misleading content* (0.709), reinforcing the association between misleading narratives and attention-grabbing techniques.

• Community Agreement: Despite differences in content type, upvote ratios remain consistently high (>0.92), indicating a generally positive reception across categories.

The temporal trends reveal fluctuations in content generation, with peaks that may correspond to external events or shifts in user activity, suggesting that contextual factors influence content production.

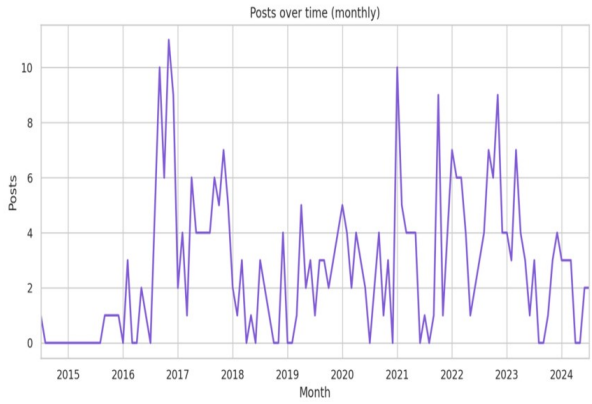

Figure 4. Posts over time

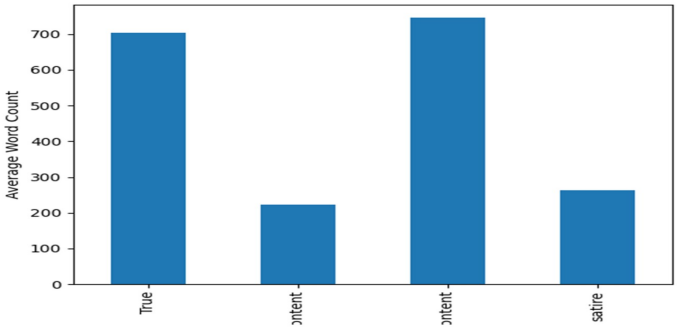

Figure 5. Average content length across content categories

The figure confirms that misleading content is significantly more verbose than other categories. This supports the hypothesis that increased length is strategically used to enhance perceived credibility and informational richness.

4.2.2 Subreddit-Level Analysis

Table 2 summarizes engagement and content characteristics at the subreddit level.

The results indicate a strong concentration of engagement within specific communities. Subreddits such as *nottheonion* and *UpliftingNews* exhibit extraordinarily high average scores (>60,000), reflecting viral content dynamics and high audience interaction.

Content strategies also vary significantly across subreddits. For instance, *neutralnews* produces the longest posts (~880 words), suggesting a detailed and information-rich approach, whereas *subsimulatorgpt2* generates notably short posts (~93 words), reflecting the concise nature of AI-generated content.

Additionally, the absence of engagement metrics in subreddits such as gpt35, gpt4, and vicuna limits direct comparisons and highlights data-collection constraints.

The dominance of subreddits such as *nottheonion* suggests that content novelty and irony significantly influence user engagement dynamics.

Subreddit	Posts	Avg Score	Median Score	Avg Upvote Ratio	Avg Content Words
Waterford whispers news	488	11.873	11	0.976	286.891
savedyouaclick	392	6286.036	5182.5	0.966	755.571
neutralnews	252	469.333	383	0.952	880.079
nottheonion	189	73168.349	72401	0.889	554.974
theonion	183	953.027	1063	0.985	247.355
subsimulatorgpt2	176	1404.881	1200.5	0.987	92.960
UpliftingNews	167	63039.497	64182	0.898	623.012
gpt35	125	—	—	—	302.168
gpt4	125	—	—	—	302.752
vicuna	112	—	—	—	246.473
satire	82	18.195	16	0.951	212.427
usanews	53	79.283	1	0.998	730.717

Table 2. Subreddit-wise Summary Statistics

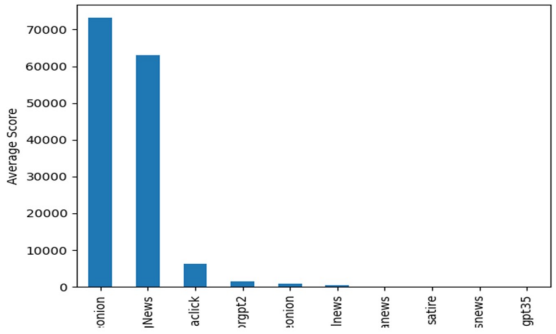

Figure 6. Average engagement scores across subreddits

The figure clearly shows that a small number of subreddits dominate overall engagement, reinforcing the unequal distribution of visibility across the platform.

4.3 Feature Engineering and Representation

To support predictive modeling, a comprehensive feature engineering process was conducted, resulting in a feature matrix of 2,344 instances and 86 features. These features were designed to capture multiple dimensions

of textual data, including structure, sentiment, and stylistic patterns.

The engineered features include:

- **Linguistic attributes** (e.g., sentence length, word length, part-of-speech ratios)
- **Sentiment and emotion features** (e.g., polarity, subjectivity, anger, joy, fear)
- **Stylistic markers** (e.g., capitalization ratio, punctuation usage)
- **Clickbait indicators**
- **TF-IDF representations**
- **Metadata and categorical encodings**

These features were selected to provide a comprehensive representation of content characteristics, enabling effective differentiation between authentic and misleading information.

Feature	Mean	Std	Min	Max
avg_sentence_length	8.7000	1.7670	5.0000	11.0000
avg_word_length	4.7030	1.0275	3.8889	7.2000
noun_ratio	0.2262	0.1615	0.0000	0.5000
verb_ratio	0.1640	0.1114	0.0000	0.4000
adj_ratio	0.1753	0.0870	0.0000	0.3000
polarity	0.0500	0.4718	-0.7500	0.7500
subjectivity	0.5833	0.3199	0.0000	1.0000
anger	0.2000	0.6325	0.0000	2.0000
fear	0.2000	0.6325	0.0000	2.0000
joy	0.3000	0.6749	0.0000	2.0000
caps_ratio	0.0600	0.1075	0.0000	0.3000
exclam_count	0.2000	0.4216	0.0000	1.0000
question_count	0.1000	0.3162	0.0000	1.0000
clickbait_flag	0.4000	0.5164	0.0000	1.0000

Table 3. Feature Summary Statistics

The feature analysis (Table 3) indicates moderate linguistic complexity, with an average sentence length of approximately 8.7 words. Subjectivity levels are relatively high **(**0.58), suggesting that many posts incorporate opinionated or interpretive language. Emotional features exhibit low-to-moderate intensity, indicating balanced emotional expression across the dataset. The presence of clickbait indicators (mean = 0.4) confirms their widespread usage.

Histograms illustrating variability in sentence length, polarity, and subjectivity across posts.

Figure 7 highlights the most frequently occurring words across titles and content.

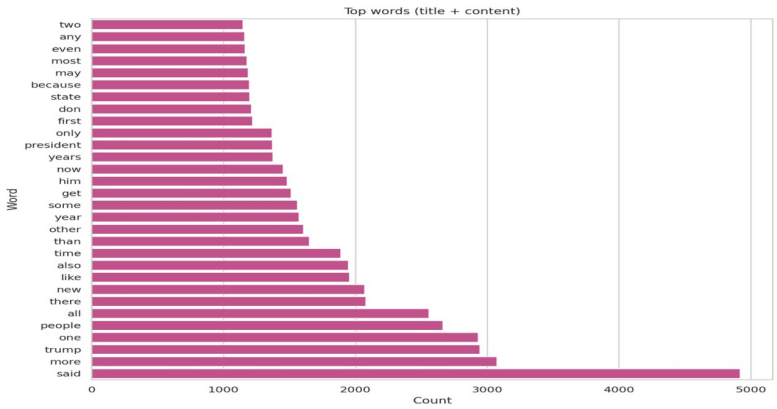

Figure 7. Top words (Title + content)

The figure reveals dominant lexical patterns and thematic concentrations, providing insights into commonly used terms across categories.

Figure 8 examines the relationship between title length and content length.

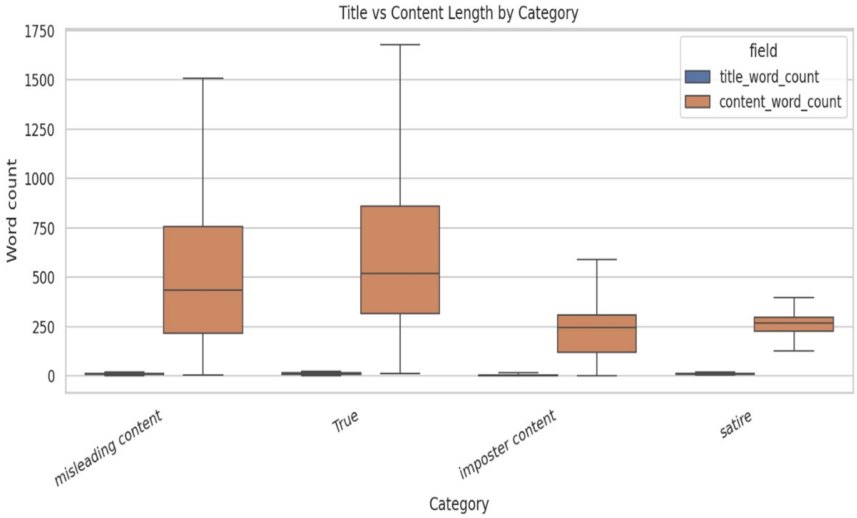

Figure 8. Title Vs Content Length by Category

The analysis shows that misleading content tends to combine relatively longer bodies with moderately sized titles, indicating a structural strategy for engagement and persuasion.

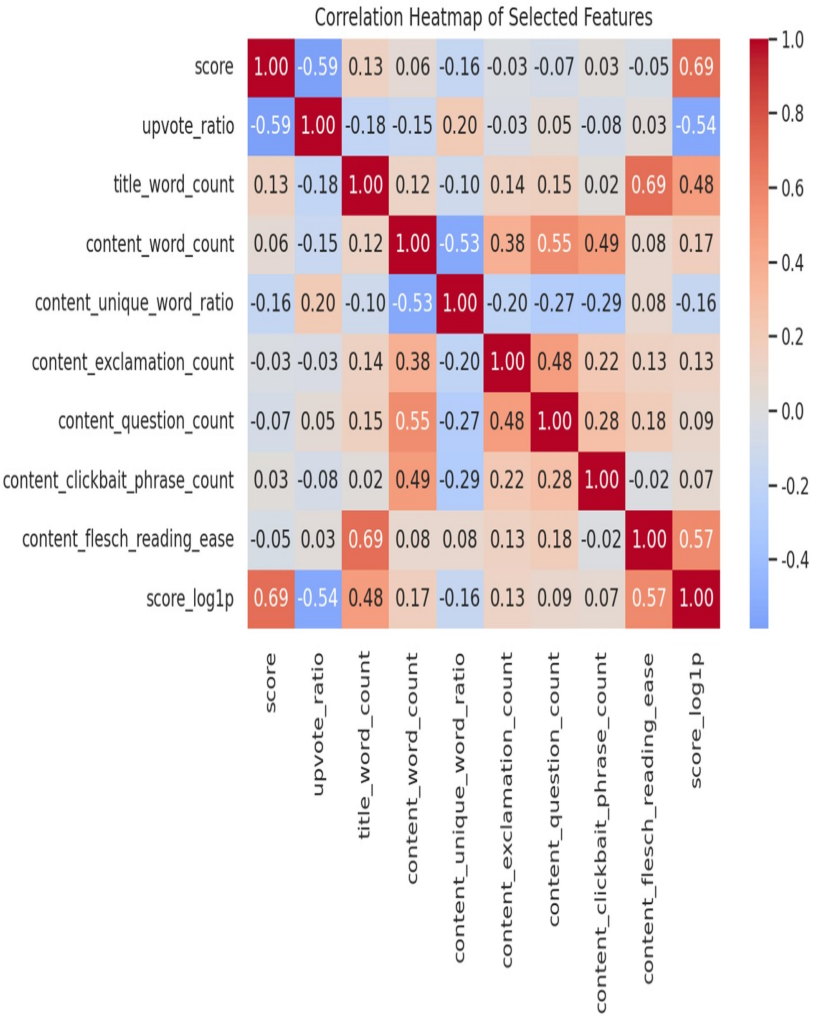

Figure 9. Correlation Map of Selected Features

Figure 9 shows correlations among linguistic and stylistic features, indicating weak-to-moderate relationships, suggesting that features contribute complementary rather than redundant information.

• The correlation matrix indicates primarily weak-to-moderate relationships, suggesting that features contribute complementary information rather than redundancy.The presence of clickbait indicators (mean = 0.4) confirms their widespread use across categories.

4.4 Predictive Modeling

4.4.1 Model Configuration

A baseline classification model was developed to evaluate the distinguishability of content categories. The model utilizes:

• TF-IDF vectorization (max_features = 3000)

• Linear Support Vector Machine (LinearSVC)

• Train-test split: 80/20

This configuration ensures computational efficiency and strong performance in high-dimensional text classification tasks.

4.4.2 Model Performance

The model demonstrates strong overall performance:

• Accuracy: 0.8870

• Precision (weighted): 0.8857

• Recall (weighted): 0.8870

• F1-score (weighted): 0.8858

Class-wise performance indicates variation across categories, with *satire* and *imposter content* achieving the highest scores, while *misleading content* remains the most challenging to classify. (Table 4).

Class	Precision	Recall	F1-score
True	0.88	0.91	0.89
Imposter Content	0.94	0.90	0.92
Misleading Content	0.76	0.71	0.73
Satire	0.92	0.96	0.94

Table 4. Class-wise Performance

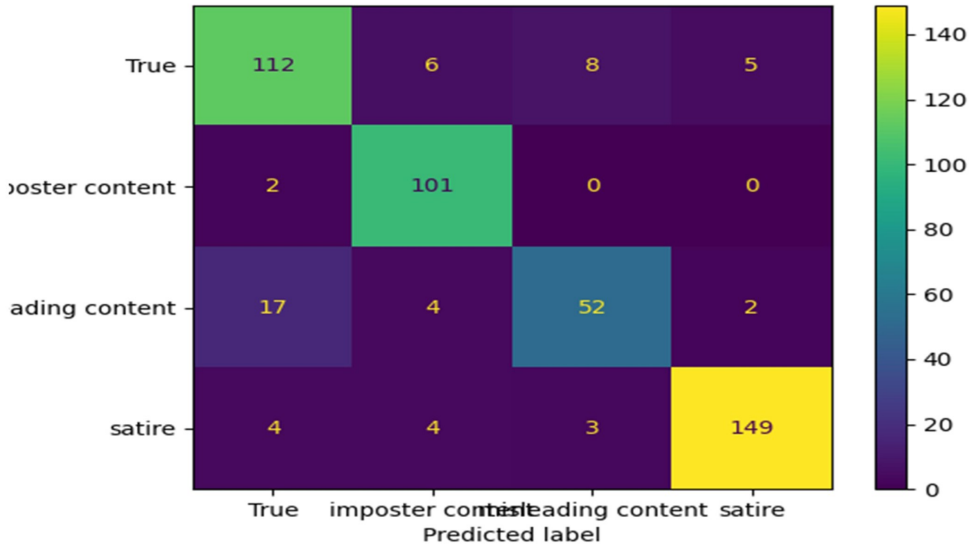

Figure 10. Confusion matrix of the classification model

The matrix reveals that most misclassifications occur between *misleading* and *true* content, indicating significant semantic overlap and highlighting the complexity of misinformation detection. Misclassification of misleading content as true suggests lexical and structural similarity, indicating that deceptive content is intentionally crafted to mimic legitimate narratives.

The matrix highlights classification accuracy across categories, with notable confusion between misleading and true content.

4.4.3 Interpretation of Results

The results demonstrate that the model effectively captures distinguishing patterns across content categories, achieving high overall accuracy (~88.7%). Categories such as satire and imposter content are more easily identifiable due to distinctive lexical and stylistic features.

However, misleading content exhibits lower recall (0.71), suggesting that it closely mimics the structure and vocabulary of true content. This underscores the inherent difficulty of detecting misinformation when deceptive content is designed to resemble legitimate narratives.

Figure 11 illustrates the most discriminative features.

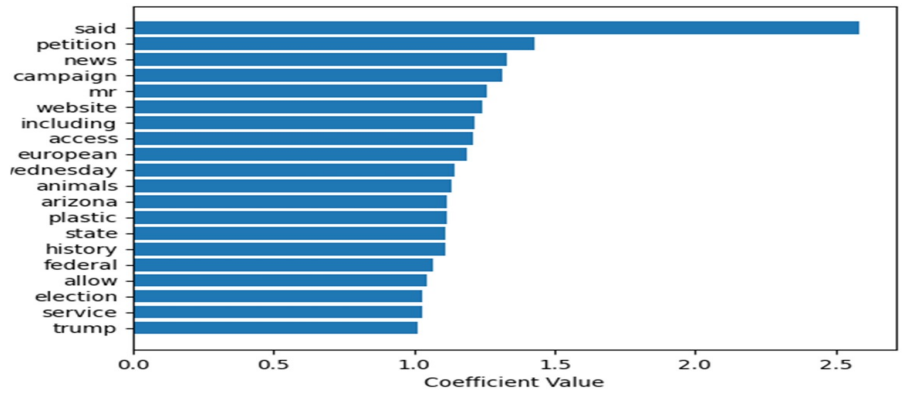

Figure 11. Top discriminative features from the TF-IDF model

The figure provides interpretability by identifying key terms that contribute most strongly to classification decisions, offering insights into how lexical patterns differentiate content types.

The figure shows the most influential terms contributing to classification across categories.

Variable	Test	p-value	Significance
Score	ANOVA	<0.05	Significant
Content Length	ANOVA	<0.05	Significant
Clickbait	Kruskal-Wallis	<0.05	Significant

Table 5. Statistical Test Results

Table 5 presents the results of inferential statistical tests assessing whether the observed differences across content categories are statistically significant. The table summarizes the outcomes of hypothesis testing for three key variables: engagement score, content length, and clickbait usage.

The results indicate that all three variables exhibit statistically significant differences across content categories ($p < 0.05$). Specifically, the analysis of engagement scores using one-way ANOVA confirms that user interaction levels vary meaningfully between categories. This finding supports earlier observations that authentic and misleading content attract different levels of attention, with certain categories benefiting from higher visibility and virality.

Similarly, the statistical test for content length reveals significant variation across categories, validating the descriptive finding that misleading content tends to be more verbose. This suggests that increased textual length may be strategically employed to enhance perceived credibility and informational depth.

The analysis of clickbait usage, assessed using a non-parametric test due to distributional considerations, also shows significant differences across categories. The results confirm that misleading content has a higher prevalence of clickbait indicators than other categories, reinforcing its role as a deliberate engagement strategy.

Overall, Table 5 provides strong empirical support for the study's central claims by demonstrating that the differences observed in descriptive analysis are not due to random variation but reflect systematic and meaningful distinctions in content structure and engagement behavior. These findings strengthen the validity of the proposed analytical framework and highlight the importance of combining descriptive and inferential approaches in misinformation research.

4.5 Key Insights and Implications

The analysis yields several important insights:

1. Engagement is not aligned with authenticity: High engagement is observed across both true and certain satirical communities, indicating that virality is driven more by content appeal than factual accuracy.

2. Misleading content mimics credibility: Increased verbosity and higher clickbait usage suggest strategic design to resemble authentic news.

3. AI-generated content differs structurally: Shorter length and missing engagement metrics distinguish AI-generated posts from human-curated content.

4. Baseline models are effective but limited: While achieving strong performance, traditional models struggle with nuanced distinctions, particularly in detecting misleading content.

This finding aligns with prior studies (Zhou and Zafarani 2020), which highlight the role of emotional and exaggerated content in driving engagement.

4.6 Limitations and Future Directions

This study is subject to several limitations. Missing engagement data in AI-generated subreddits restricts comprehensive comparison across all categories. Additionally, dataset imbalance and subreddit-specific biases may influence both descriptive and predictive outcomes.

Future research can address these limitations by incorporating:

- Transformer-based models (e.g., BERT) for improved semantic understanding
- Temporal modeling to capture evolving engagement dynamics
- Network-based analysis to study information propagation patterns

5. Conclusion

This study presented a comprehensive analysis of Reddit-based content to investigate the structural, linguistic, and engagement characteristics of *true*, *satirical*, *imposter*, and *misleading* information. By integrating descriptive analytics with machine learning–based classification, the study provides both empirical insights and predictive understanding of how different content types behave within online communities.

The findings demonstrate that user engagement is not necessarily aligned with content authenticity. While true content achieves the highest average engagement, certain satirical and community-driven posts also achieve high visibility, suggesting that virality is driven more by narrative appeal and audience resonance than by factual accuracy alone. At the same time, misleading content emerges as structurally distinct, characterized by greater verbosity and higher use of clickbait markers, suggesting deliberate strategies to simulate credibility and attract attention.

Subreddit-level analysis further reveals that engagement is highly concentrated within a small number of communities, highlighting the role of platform dynamics in shaping content visibility. Additionally, AI-generated content shows clear structural differences, particularly in terms of shorter length and incomplete engagement metadata, which may impact its comparability with human-generated posts.

From a predictive perspective, the baseline classification model achieved strong overall performance, confirming that textual and stylistic features are effective for content categorization. However, the relatively lower performance in detecting misleading content underscores a critical challenge: deceptive information often closely mimics legitimate news in both language and structure. This highlights the limitations of traditional machine learning approaches in capturing nuanced semantic intent.

These findings underscore the need for integrated, context-aware detection systems that go beyond textual analysis to incorporate engagement and platform dynamics.

Overall, this study contributes to the growing body of research on misinformation by demonstrating that content characteristics, user engagement, and platform structures are deeply interconnected. The results emphasize the need for more advanced approaches that combine linguistic analysis with contextual and network-level signals.

Future work should focus on leveraging transformer-based models, incorporating temporal and propagation dynamics, and exploring user interaction networks to enhance detection accuracy and robustness. Such advancements are essential for developing scalable and reliable systems to combat misinformation in increasingly complex digital ecosystems.

References

[1] Zhou, X., Zafarani, R. (2020). A survey of fake news: fundamental theories, detection methods, and opportunities. *ACM Comput Surv* 53(5):1–40

[2] Baptista, J. P., Gradim, A. (2020). Understanding fake news consumption: a review. Soc Sci 9(10):185

[3] Zhou, X., Zafarani, R., Shu, K., Liu, H. (2019). Fake news: fundamental theories, detection strategies and challenges, In: Proceedings of the twelfth ACM international conference on web search and data mining, WSDM'19. *Association for Computing Machinery, New York, NY, USA*, p 836–837.

[4] Mingxiao, Sui., Ian, Hawkins., Rui, Wang. (2023). Computers in Human Behavior When falsehood wins? Varied effects of sensational elements on users' engagement with real and fake posts, *Computers in Human Behavior*. Volume 142, May 2023, 107654.

[5] Bhalerao, R., Aliapoulios, M., Shumailov, I., Afroz, S., McCoy, D. (2019). Mapping the underground: supervised discovery of cybercrime supply chains 2019 *APWG Symposium on Electronic Crime Research (eCrime)* p. 1-16.

[6] Stringhini, G., Egele, M., Kruegel, C. (2012). Vigna Poultry markets: on the underground economy of twitter followers Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks, *Association for Computing Machinery, New York, NY, USA* (2012), p. 1-6.

[7] Qi, P., Cao, J., Yang, T., Guo, J, L i, J. (2019). Exploiting multi-domain visual information for fake news detection. *In: IEEE international conference on data mining* (ICDM), p 518–527.

[8] Lang, P. J. (1979) A bio-informational theory of emotional imagery. *Psychophysiology* 16(6):495–512.

[9] Qian, Xiao., Elisa, Bertino. (2017). Detecting deceptive engagement in social media by temporal pattern analysis of user behaviors: a survey, WIRes, Volume 7 (5). p. e1210.

[10] Kuo, H. Y., Chen, S, Y. (2025). Predicting User Engagement in Health Misinformation Correction on Social Media Platforms in Taiwan: *Content Analysis and Text Mining Study, J Med Internet* Res 2025;27:e65631

[11] Ozan, Candogan., Kimon, Drakopoulos. (2020). Optimal Signaling of Content Accuracy: Engagement vs. Misinformation, *Operations Research* Vol. 68 (2)

[12] Maria, Castaldo. (2022). Attention dynamics on YouTube: conceptual models, temporal analysis of engagement metrics, fake views. Automatic. Université Grenoble Alpes [2020], *Thesis submitted to Universit'e Grenoble Alpes*.

[13] Van der Meer, T G., Jin, Y. (2020). Seeking formula for misinformation treatment in public health crises: the effects of corrective information type and source. *Health Commun*. May 35(5) 560-575.

[14] Zeng, H. K., Lo, S, Y., L i S S. (2024). Credibility of misinformation source moderates the effectiveness of corrective messages on social media. *Public Underst Sci*. Jul 31, 33(5):587-603.

[15] Martel, C., Mosleh, M., Rand, D G. (2021). You're definitely wrong, maybe: correction style has minimal effect on corrections of misinformation online. *Media Commun*. Feb 03, 9(1):120-133.

[16] Bode, L., Vraga, E K. (2018). See something, say something: correction of global health misinformation on social media. *Health Commun*. Sep 2018; 33 (9):1131-1140.

[17] Wang, Y. (2021). Debunking misinformation about genetically modified food safety on social media: can heuristic cues mitigate biased assimilation? *Sci Commun*. *Jun* 18, 43(4):460-485.

[18] Song, Y., Wang, S., Xu, Q. (2022). Fighting misinformation on social media: effects of evidence type and presentation mode. *Health Educ Res*. May 24, 37(3):185-198

[19] Sui, Y., Zhang, B. (2021). Determinants of the perceived credibility of rebuttals concerning heal th misinformation. *Int J Environ Res Public Health*. Feb 02, 2021;18(3):1345.

[20] Uppada, S. K., Manasa, K., Vidhathri, B. et al. (2022). Novel approaches to fake news and fake account detection in OSNs: *user social engagement and visual content centric model. Soc.* Netw. Anal. Min. 12, 52 .

[21] Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., Liu, H. (2019). Unsupervised fake news detection on social media: a generative approach. In: Proceedings of the AAAI conference on artificial intelligence, vol 33,(01), p 5644–5651.

[22] Liu, Y., Wu, Y B. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18), *New Orleans, Louisiana,* USA, February 2-7, 2018. AAAI Press, p 354–361.

[23] Dordevic, M., Pourghomi, P., Safieddine, F. (2020) Identifying fake news from the variables that governs the spread of fake news. *In:15th international workshop on semantic and social media adaptation and personalization (SMA). IEEE*, p 1–6.

[24] Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. *In: Proceedings of the 25th ACM international conference on Multimedia,* p 795–816.

[25] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. *In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining,* p 849–857.

[26] Khalil, H. A., Maged, S. A. (2021). Deepfakes creation and detection using deep learning. In: 2021

International mobile, intelligent, and ubiquitous computing conference (MIUCC). *IEEE,* p 1–4.

[27] Boshmaf, Y., Logothetis, D., Siganos, G., Lería, J., Lorenzo, J., Ripeanu, M., Beznosov, K., Halawa, H . (2016). Integro: Leveraging victim prediction for robust fake account detection in large scale OSNs. *Comput Secur* 61:142–168

[28] Wang, G., Konolige, T., Wilson, C., Wang, X., Zheng, H., Zhao, B. Y. (2013). You are how you click: clickstream analysis for sybil detection. *In:Proceedings of the 22nd USENIX conference on security, SEC'13. USENIX Association, USA,* p 241–256.

[29] Viswanath, B., Bashir, M. A., Crovella, M., Guha, S., Gummadi, K. P., Krishnamurthy, B., Mislove, A. (2014). Towards detecting anomalous user behavior in online social networks. *In: Proceedings of the 23rd USENIX conference on security symposium, SEC'14. USENIX Association, USA*, p 223–238

[30] Akyon, F C., Kalfaoglu, E. (2019). Instagram fake and automated account detection. *In: Innovations in intelligent systems and applications conference (ASYU)*, p 1–7.

[31] Khaled, S., El-Tazi, N., Mokhtar, H. M. O. (2018). Detecting fake accounts on social media. *In:IEEE international conference on Big Data (Big Data),* p 3672–3681

[32] Galan-Garcia, P., de la Puerta, J. G., Gomez, C L., Santos, I., Garcia Bringas, P. (2016). Supervised machine learning for the detection of troll profiles in Twitter social network: *application to a real case of cyber bullying*. *Log J IGPL* 24(1):42–53

[33] Budak, C., Garrett, R K., Resnick, P., Kamin, J. (2017). Threading is sticky: how threaded conversations promote comment system user retention. *Proc ACM Hum-Comput Interact*. 1:1–20.

[34] Eveland, J. r., W. P. (2003). A "mix of attributes" approach to the study of media effects and new communication technologies. *J Commun*. 53(3):395–410.

[35] Friess, D., Eilders, C. (2015). A systematic review of online deliberation research. *Policy Internet*. 7(3):319–339.

[36] Robert, M., Bond, R., Kelly Garrett. (2023). Engagement with fact-checked posts on Reddit, PNAS Nexus, Volume 2 (3) March 2023, pgad018.

[37] Raja, desingan., A., Resnick, P., Budak, C. (2017). Quick, community-specific learning: how distinctive toxicity norms are maintained in political subreddits. Proc Int AAAI Conf Web Soci Media. 14(1) 557–568.

[38] Chandrasekharan, E., et al. 2018. The internet's hidden rules: an empirical study of Reddit norm violations at micro, meso, and macro scales. *Proc ACM Hum-Comput Interact*. 2:1–25.

[39] Shearer, E., Mitchell, A. (2022). News use across social media platforms in 2020. Technical report, Pew Research Center, Washington (DC), January 1.

[40] Horne, B. D., Adali, S., Sikdar, S. (2017). Identifying the Social Signals That Drive Online Discussions: A Case Study of Reddit Communities,"26th *International Conference on Computer Communication and Networks (ICCCN), Vancouver, BC, Canada*, p. 1-9.

[41] Yulin, Y. u., Julie, Jiang., Paramveer, S., Dhillon. (2024). Characterizing the Structure of Online Conversations Across Reddit, *Proceedings of the ACM on Human-Computer Interaction*, Volume 8, Issue CSCW2. Article No.: 374, Pages 1 -23.