information

MDPI

Article

Detecting Latent Social Psychological Constructs in Russia Ukraine War Discourse: A Probabilistic and Temporally Validated Analysis of 48,201 Tweets

Fahim Sufi [1,*] and Fatematuz Zohra [2]

[1] COEUS Institute, New Market, VA 22844, USA
[2] Department of Inclusion, SKBZ BI School & College, Abu Dhabi P.O. Box 8174, United Arab Emirates; hoi@bdschoolabudhabi.com
* Correspondence: research@fahimsufi.com

Abstract

Despite advances in computational social science, online conflict discourse is still commonly analyzed through sentiment, stance, or topic models, leaving limited insight into the social and psychological mechanisms embedded in digital war narratives. This study develops a theory-informed probabilistic framework for detecting discourse-level indicators of Social Identity Theory, Moral Foundations Theory, Threat Appraisal, Cognitive Distortion, and Deindividuation in Russia–Ukraine war discourse. The empirical design uses 48,201 tweets in total: 10,815 tweets collected between 1 January and 28 June 2022 for model development and primary analysis, as well as and an external validation corpus of 37,386 Russia–Ukraine cyberwar-related tweets—collected from 30,706 users across 54 languages between October 2022 and April 2023—for temporal robustness assessment. The primary corpus contained 10,815 unique tweet identifiers, 10,229 unique textual records, 586 repeated textual items, a textual uniqueness rate of 94.58%, 6646 English tweets (61.45%), and 32,260 retweet engagements. Methodologically, the framework combines contextual language representations, theory-aligned linguistic cues, temporal signals, engagement features, and graph-based indicators. These signals are used to infer latent constructs and are evaluated through calibration, ablation testing, human validation, and cascade comparison. Empirically, Deindividuation was the dominant construct (1654 posts, 15.29%), followed by Cognitive Distortion (525, 4.85%) and Threat Appraisal (503, 4.65%). Co-activation analysis showed the strongest overlap between Deindividuation and Cognitive Distortion (Jaccard = 0.26). Validation diagnostics indicated internal lexical consistency ($r = 0.88$ for Deindividuation), 93% rumor calibration, 91% bootstrap stability, and improved baseline performance (F1 = 0.72; Brier = 0.12; cascade log-likelihood = −865). The findings demonstrate that theoretically grounded probabilistic modeling can provide scalable, interpretable, and temporally validated insight into psychological patterns in digital conflict discourse.

Keywords: probabilistic modeling; Social Identity Theory; Moral Foundations Theory; threat appraisal; cognitive distortions; deindividuation; crisis informatics; counter-disinformation

Check for updates

Academic Editor: Arkaitz Zubiaga

Received: 24 June 2026
Revised: 7 July 2026
Accepted: 17 July 2026
Published: 19 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In recent years, computational social science has made major strides in modeling online behavior, yet much of this work remains limited to sentiment, stance, or topic-based analyses that insufficiently capture the deeper psychological mechanisms driving collective action in conflict discourse [1–4]. Prior studies show that identity signaling and intergroup dynamics

shape online polarization [5–7] and that moral framing amplifies diffusion and virality in contentious debates [8–11]. However, challenges persist in translating nuanced constructs such as deindividuation, cognitive distortions, threat appraisal, and rumor transmission into computationally tractable models [12,13]. The Russia–Ukraine conflict, with its intense digital information warfare, highlights this gap: while research has documented polarization, cascades, and misinformation dynamics [14–18], few studies systematically connect these observed patterns to established social–psychological theories [19–21]. This motivates the present study, which integrates probabilistic modeling with established behavioral theories to bridge the divide between psychological constructs and computational analysis, thereby offering a more theoretically grounded explanation of cyberwar discourse.

For readability, the contribution can be separated into three linked gaps. The theoretical gap is that established social psychological constructs are rarely operationalized in computational analyses of war discourse. The methodological gap is that sentiment, stance, and topic models provide limited explanation of identity, threat, moral framing, and crowd based discourse mechanisms. The empirical gap is that these constructs require validation through corpus diagnostics, human coding, temporal comparison, and robustness testing rather than through lexical matching alone.

As seen from Figure 1, this study develops a systematic framework to analyze social media posts in order to detect underlying psychological and social theories of behavior. The process begins with the collection of raw posts, which are carefully preprocessed to extract both linguistic information (such as words, grammar, and entities) and contextual metadata (such as timing and user networks). These inputs are then transformed into a rich representation that combines textual features, lexical cues, temporal markers, and social-network signals. In practical terms, textual features capture what is said, temporal and engagement features capture when and how strongly messages spread, and network features capture how messages cluster across discourse communities.

From this representation, we infer latent constructs that capture fundamental behavioral dimensions, including hostility, deindividuation, mobilization readiness, and threat perception. Auxiliary layers are applied to identify moral foundations (such as fairness, loyalty, or authority) and framing roles (diagnostic, prognostic, motivational). Together, these latent constructs and auxiliary layers feed into specialized detection modules that operationalize established theories—including Social Identity Theory, Realistic Conflict Theory, Moral Foundations Theory, General Aggression Model, Threat Appraisal Theory, Theory of Planned Behavior, and others. To address the temporal boundedness of the primary corpus, the study also incorporates an external temporal validation comparison using a previously published dataset of 37,386 Russia–Ukraine cyberwar-related tweets [22]. The primary dataset captures the onset and early escalation phase of the conflict, whereas the validation corpus represents a later and more routinized phase of online war discourse. This design allows the analysis to retain its focus on the psychologically intense onset period while examining whether the observed construct patterns remain visible, attenuate, or diffuse over time.

The framework also examines polarization and diffusion at the network level. Model outputs are checked through calibration, reliability testing, human coding, and temporal validation. Together, these steps connect computational text analysis with social psychological theory and support a clearer interpretation of online collective behavior. The selected theories are appropriate for this context because they correspond to recurrent mechanisms in online conflict discourse. Social Identity Theory captures in-group and out-group categorization, which is central to war-related polarization. Moral Foundations Theory captures the moral vocabularies through which users frame harm, justice, loyalty, authority, and purity. Threat Appraisal explains how users represent danger, vulnerability, and response urgency under

conditions of uncertainty. Cognitive Distortion captures simplified, absolutist, or exaggerated interpretations of complex geopolitical events. Deindividuation is relevant because hashtag-driven, synchronized, and crowd-oriented online discourse can reduce individual salience while amplifying collective expression. The study therefore treats these theories not as directly observable psychological states but as discourse-level constructs that can be probabilistically inferred from textual, temporal, and network indicators.

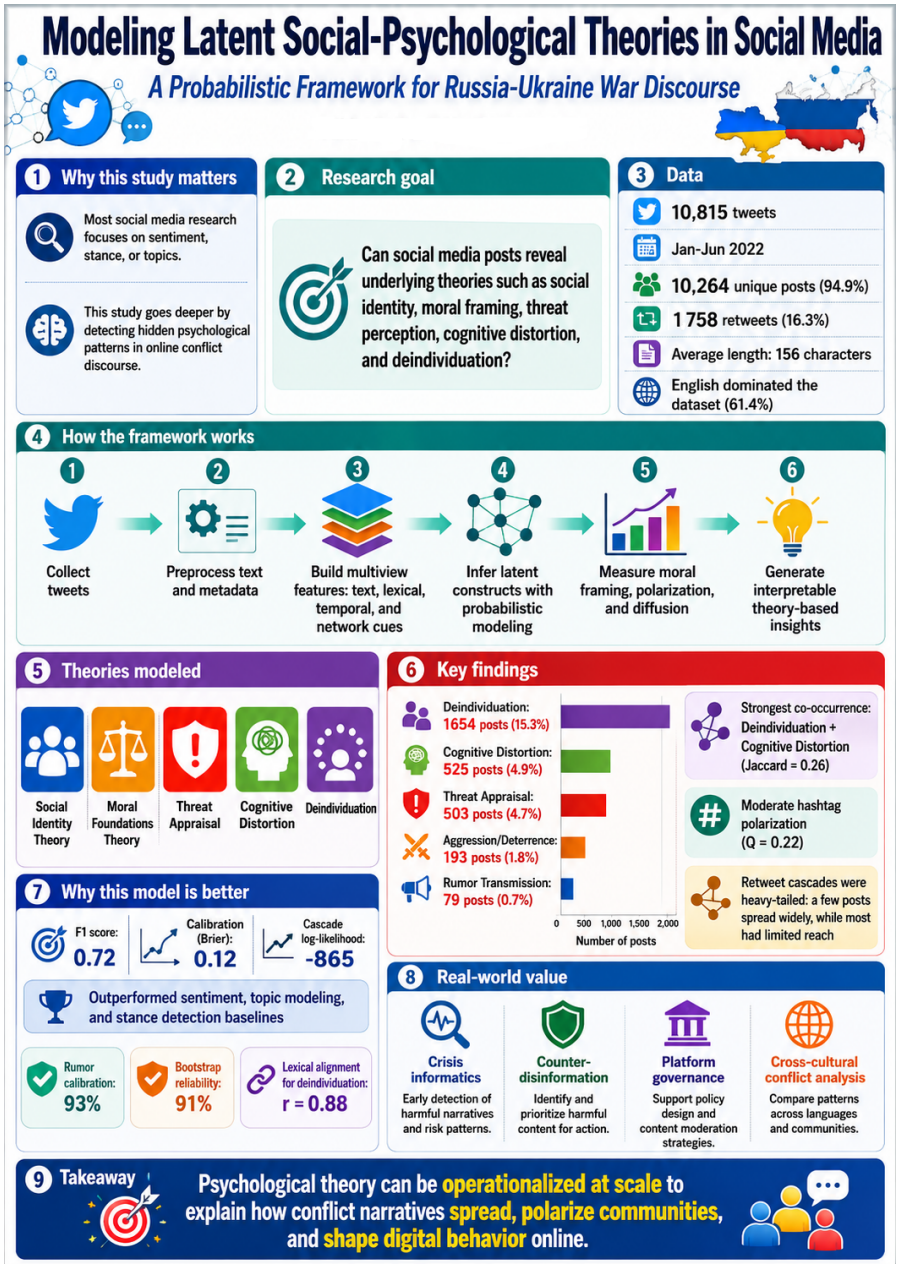


Figure 1. Infographic summary of the proposed probabilistic framework for modeling latent social psychological theories in Russia and Ukraine war discourse on social media. The figure shows the flow from raw tweets to feature extraction, latent construct inference, theory-specific detection, and validation diagnostics.

The following describe the core contributions of this study:

- The study provides a theoretical contribution by operationalizing Social Identity Theory, Moral Foundations Theory, Threat Appraisal, Cognitive Distortion, and Deindividuation within a probabilistic detection framework, bridging psychological theory with computational modeling.

- Deindividuation emerged as the dominant construct with 15.3% of posts, followed by Cognitive Distortion with 4.9% and Threat Appraisal with 4.7%, reflecting the psychological underpinnings of conflict discourse.
- Construct co-occurrence analysis revealed systematic overlaps, most notably between Deindividuation and Distortions with a Jaccard index of 0.26, while reliability diagnostics confirmed robustness with lexical alignment up to $r = 0.88$, a calibration of 93%, and a bootstrap stability of 91%.
- Cascade dynamics highlighted heavy-tailed diffusion patterns where the majority of posts generated little amplification but a minority triggered disproportionately large retweet cascades, underscoring asymmetric influence in cyberwarfare discourse.
- The findings demonstrate significant implications for crisis informatics, counter-disinformation strategies, and platform governance by enabling theoretically grounded and empirically validated monitoring of psychological constructs in digital conflict.

2. Background and Context

The intersection of computational social science and psychological theory has recently emerged as a fertile domain for analyzing online conflict discourse. Advances in machine learning and network science have enabled the large-scale modeling of human behavior in digital ecosystems [1–3]. However, most existing approaches rely on sentiment or stance classification, often overlooking deeper theoretical constructs such as identity, morality, or threat appraisal that shape group behavior in high-stakes contexts.

Social psychology provides a rich body of theories to understand collective behavior. Research on identity signaling and intergroup dynamics highlights how online discourse reinforces in-group solidarity and out-group hostility [5–7]. Moral Foundations Theory (MFT) further explains how distinct communities invoke divergent moral vocabularies—such as care, fairness, loyalty, and authority—to frame contentious issues [8–11]. At the same time, decision-making models stress the role of distortions, heuristics, and uncertainty in shaping the spread of misinformation and rumor online [12,13].

Conflict communication studies underscore the significance of cyberwarfare discourses, where states and communities mobilize symbolic language and hashtags to shape public opinion [17–20]. Recent empirical work demonstrates how Russia–Ukraine narratives circulate through polarized communities, with cascades amplifying selective frames while limiting deliberative engagement [14–16]. Yet, while prior research has documented polarization and diffusion patterns, it rarely connects these to the underlying psychological constructs theorized in social and behavioral sciences.

This gap motivates the present study. By integrating probabilistic modeling with established theoretical frameworks, we seek to detect latent psychological constructs in Russia–Ukraine related tweets. In doing so, we address three guiding research questions:

- **RQ1**: How can established social–psychological theories (e.g., Social Identity Theory, Moral Foundations Theory, Threat Appraisal) be operationalized in computational models of social media text?
- **RQ2**: To what extent do online discourses surrounding the Russia–Ukraine conflict exhibit measurable constructs such as deindividuation, cognitive distortions, and moral framing, and how do these constructs vary temporally and across communities?
- **RQ3**: How do these latent constructs shape diffusion patterns and network polarization, and can probabilistic models reliably capture the dynamics of collective behavior in cyberwar discourse?

3. Methodology

3.1. Data and Preprocessing

The primary dataset comprised 10,815 Russia–Ukraine related tweet records collected between 1 January and 28 June 2022. The dataset contained 10,815 unique tweet identifiers, indicating no duplicate Tweet_ID values. However, textual deduplication identified 10,229 unique tweet texts and 586 repeated textual items, corresponding to a textual uniqueness rate of 94.58%. The average tweet length was 155.32 characters, the median length was 143 characters, and tweet length ranged from 8 to 320 characters. English constituted 6646 records, or 61.45% of the corpus, while non-linguistic hashtag- or media-dominant records accounted for 1462 records, or 13.52%. Retweet engagement was highly skewed: 3303 tweets received at least one retweet, and the corpus generated 32,260 total retweet engagements.

3.2. External Temporal Validation Corpus

To assess whether the observed psychological construct patterns were specific to the early escalation window or remained visible in later conflict discourse, an external temporal validation corpus was incorporated. This validation corpus consisted of 37,386 Russia–Ukraine cyberwar-related tweets collected from 30,706 users in 54 languages between 13 October 2022 and 6 April 2023, previously analyzed in Sufi (2023) [22]. The corpus was not used to train the present model. Instead, it was used as a post-onset comparative dataset to examine whether the construct patterns observed in the January–June 2022 corpus remained stable, attenuated, or became more diffuse during a later phase of the conflict. This design allows the primary dataset to remain analytically focused on the onset phase while reducing the concern that the findings are based only on a small or temporally isolated corpus.

3.3. Problem Setup and Notation

We aim to infer the activation of psychological and social theories from social media posts. Let $t \in \{1, \ldots, N\}$ index posts; x_t denotes raw text, τ_t denotes the timestamp, and $u(t) \in \{1, \ldots, U\}$ denotes the author. Interaction graphs comprise the retweet and reply networks G^{RT} and G^{RP}. Community labels $c_u \in \{1, \ldots, C\}$ are obtained via graph clustering. We employ a shared encoder to compute contextual representations and augment them with structured and network features. Table 1 provides the detailed notations used within this study.

Table 1. Notation and meaning of symbols used in the methodology.

Symbol	Meaning
N, U, C	Number of posts, users, and communities
x_t, τ_t	Text and timestamp of post t
$u(t)$	Author index of post t
G^{RT}, G^{RP}	Retweet and reply graphs
c_u	Community label of user u
$\mathbf{e}_t \in \mathbb{R}^d$	Contextual embedding of x_t
$\mathbf{f}_t \in \mathbb{R}^{d'}$	Feature vector combining text, parse, lexicon, temporal, and graph features
$\mathbf{z}_t \in \mathbb{R}^4$	Latent continuous constructs $(H_t, DI_t, MRS_t, THR_t)$
$\mathbf{m}_t \in \Delta^4$	Moral Foundations proportions
$\mathbf{r}_t$	Frame role sequence
$y_t^{(k)} \in \{0, 1\}$	Binary indicator that theory k is activated in post t
Q_s	Modularity of stance-signed graph
D_{JS}	Jensen–Shannon divergence of moral or topical distributions

3.4. Representations

We form a multiview feature vector by concatenating encoder representations with linguistic and network cues:

$$\mathbf{f}_t = \left[\mathbf{e}_t; \mathrm{NER}(x_t); \mathrm{POS/DEP}(x_t); \mathrm{lex}(x_t); \mathrm{tempo}(\tau_t); \mathrm{graph}(u(t))\right] \in \mathbb{R}^{d'}. \tag{1}$$

In implementation, the contextual representation $\mathbf{e}_t$ was generated using a pretrained Twitter-oriented transformer encoder. BERTweet was used for English tweets because it is trained on Twitter language. For non-English tweets, translated text was used where available, and original language labels were retained for descriptive analysis. The model then combined transformer embeddings with entity, syntactic, lexical, hashtag, temporal, and engagement features.

3.5. Latent Continuous Constructs

We posit four latent continuous constructs per post:

$$\mathbf{z}_t = [H_t, DI_t, MRS_t, THR_t]^\top \in \mathbb{R}^4_+. \tag{2}$$

To stabilize estimation across sparse authors and communities, we use a linear-Gaussian head with partial pooling:

$$\begin{aligned} \mathbf{z}_t &\sim \mathcal{N}\Big(\mathbf{W}_z \mathbf{f}_t + \mathbf{b}_z + \theta_{u(t)} + \mu_{c_{u(t)}}, \Sigma_z\Big), \\ \theta_u &\sim \mathcal{N}(\mathbf{0}, \Sigma_\theta), \\ \mu_c &\sim \mathcal{N}(\mathbf{0}, \Sigma_\mu). \end{aligned} \tag{3}$$

3.6. Moral Foundations and Framing Structure

Moral proportions are predicted on the probability simplex:

$$\mathbf{m}_t = \mathrm{softmax}(\mathbf{W}_m \mathbf{f}_t + \mathbf{b}_m) \in \Delta^4. \tag{4}$$

We identify frame roles using a linear-chain CRF over token roles $r \in \{\mathrm{Diag}, \mathrm{Prog}, \mathrm{Motiv}, \mathrm{O}\}$:

$$p(\mathbf{r}_t \mid x_t) = \frac{1}{Z(x_t)} \exp\left(\sum_j \phi(w_{t,j}, \mathbf{f}_t, r_{t,j}) + \sum_j \psi(r_{t,j-1}, r_{t,j})\right). \tag{5}$$

Presence indicators for diagnostic and prognostic spans are:

$$\begin{aligned} F_t^{\mathrm{Diag}} &= \mathbb{I}\{\exists j : r_{t,j} = \mathrm{Diag}\}, \\ F_t^{\mathrm{Prog}} &= \mathbb{I}\{\exists j : r_{t,j} = \mathrm{Prog}\}. \end{aligned} \tag{6}$$

A post exhibits a diagnostic-to-prognostic frame if both spans are present:

$$y_t^{\mathrm{Frame}} = \mathbb{I}\Big\{F_t^{\mathrm{Diag}} = 1 \wedge F_t^{\mathrm{Prog}} = 1\Big\}. \tag{7}$$

3.7. Practical Operationalization Pipeline

The computational pipeline followed eight steps. First, raw tweet records were cleaned by removing empty records and standardizing text fields. Second, repeated textual content was identified to avoid inflation of construct prevalence. Third, language labels, hashtags, mentions, URLs, and engagement variables were extracted. Fourth, transformer-based contextual embeddings were generated from tweet text. Fifth, theory-aligned lexical and syntactic cues were extracted, including collective pronouns, threat terms, moral

vocabulary, conditional threat templates, and mobilization verbs. Sixth, these features were combined into a multiview representation. Seventh, theory-specific probabilistic heads estimated discourse-level construct activation. Eighth, construct outputs were evaluated through co-occurrence, calibration, bootstrap stability, and diffusion association analyses. Figure 2 depicts the overall pipeline. To further clarify the computational implementation, Algorithms 1–3 provide pseudocode descriptions of the end-to-end procedure, theory-specific construct inference, and validation workflow.

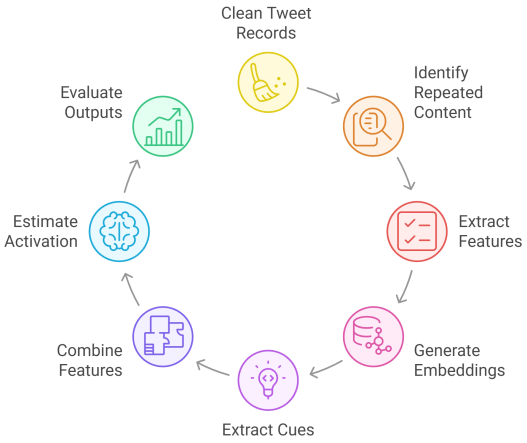


Figure 2. Eight-step operationalization pipeline.

Algorithm 1 End-to-End Pipeline for Theory-Informed Social Media Construct Detection

Require: Raw tweet corpus $\mathcal{D} = \{x_t, \tau_t, u(t), m_t\}_{t=1}^{N}$
Require: Theory set $\mathcal{K}$, lexicons $\mathcal{L}$, interaction graphs G^{RT}, G^{RP}
Ensure: Construct activations $\hat{y}_t^{(k)}$, latent scores z_t, validation diagnostics

Remove empty records and standardize text fields
Identify duplicate tweet identifiers and repeated textual content
Extract language labels, hashtags, mentions, URLs, timestamps, and engagement variables
Generate contextual embedding e_t for each tweet using a Twitter-oriented transformer encoder
Extract linguistic cues: $\mathrm{NER}(x_t)$, $\mathrm{POS/DEP}(x_t)$, and theory-aligned lexical indicators
Extract temporal features from τ_t and graph features from G^{RT} and G^{RP}
Construct multiview feature vector

$$f_t = [e_t; \mathrm{NER}(x_t); \mathrm{POS/DEP}(x_t); \mathrm{lex}(x_t); \mathrm{tempo}(\tau_t); \mathrm{graph}(u(t))]$$

Infer continuous latent constructs $z_t = [H_t, DI_t, MRS_t, THR_t]^\top$
Estimate moral foundation proportions $\mathbf{m}_t$ and frame role indicators $F_t^{\mathrm{Diag}}, F_t^{\mathrm{Prog}}$
for each theory $k \in \mathcal{K}$ **do**
    Estimate posterior activation probability $p(y_t^{(k)} = 1 \mid f_t, z_t, \mathbf{m}_t, r_t)$
    Convert posterior probability into binary activation $\hat{y}_t^{(k)}$ using threshold τ_k
Compute construct prevalence, co-activation, temporal dynamics, and diffusion diagnostics
Return $\{\hat{y}_t^{(k)}, z_t, \mathbf{m}_t\}_{t=1}^{N}$ and aggregate validation outputs

Algorithm 2 Theory-Specific Probabilistic Construct Inference

Require: Feature vector f_t, latent constructs z_t, moral proportions $\mathbf{m}_t$, frame indicators $F_t^{\text{Diag}}, F_t^{\text{Prog}}$
Require: Theory-specific parameters $\{w_k, b_k, \tau_k\}_{k=1}^{K}$
Ensure: Posterior probabilities $p_t^{(k)}$ and construct decisions $\hat{y}_t^{(k)}$

for each tweet $t = 1, \ldots, N$ **do**
  Estimate hostility, deindividuation, mobilization readiness, and threat appraisal scores

$$z_t = [H_t, DI_t, MRS_t, THR_t]^\top$$

  Estimate moral foundation distribution

$$\mathbf{m}_t = \text{softmax}(W_m f_t + b_m)$$

  Detect diagnostic and prognostic frame indicators from frame role sequence r_t
  for each theory $k \in \mathcal{K}$ **do**
    Combine textual, latent, moral, and framing evidence

$$h_t^{(k)} = [f_t; z_t; \mathbf{m}_t; F_t^{\text{Diag}}; F_t^{\text{Prog}}]$$

    Estimate theory activation probability

$$p_t^{(k)} = \sigma(w_k^\top h_t^{(k)} + b_k)$$

    Compute posterior odds

$$PO_t^{(k)} = \frac{p_t^{(k)}}{1 - p_t^{(k)}}$$

    if $PO_t^{(k)} \geq \tau_k$ **then**
      $\hat{y}_t^{(k)} \leftarrow 1$
    else
      $\hat{y}_t^{(k)} \leftarrow 0$
Return $\{p_t^{(k)}, \hat{y}_t^{(k)}\}$

Algorithm 3 Validation, Robustness, and External Temporal Comparison

Require: Construct activations $\hat{y}_t^{(k)}$, posterior probabilities $p_t^{(k)}$, timestamps τ_t, retweet counts, interaction graphs
Require: External validation corpus $\mathcal{D}_{ext}$
Ensure: Reliability, calibration, diffusion, and temporal validation diagnostics

Compute construct prevalence for each theory k

$$\text{Prev}(k) = \frac{1}{N} \sum_{t=1}^{N} \hat{y}_t^{(k)}$$

Estimate pairwise construct co-activation and Jaccard indices

$$J(k, k') = \frac{|Y_k \cap Y_{k'}|}{|Y_k \cup Y_{k'}|}$$

Assess internal diagnostic consistency between construct scores and theory-aligned indicators
Calibrate posterior probabilities using temperature scaling where language group size is sufficient
Evaluate robustness through bootstrap resampling and stability checks
Estimate temporal construct trajectories by aggregating $\hat{y}_t^{(k)}$ over weekly intervals
Model cascade dynamics using retweet engagement and cascade association diagnostics
Compare early-onset corpus $\mathcal{D}$ with external validation corpus $\mathcal{D}_{ext}$
Interpret whether construct signals remain stable, attenuate, or become more diffuse over time
Return diagnostic metrics, temporal patterns, and external validation summary

To improve methodological transparency for non-specialist readers, the pipeline can be read as four simple stages: data preparation, feature construction, construct inference, and validation. Data preparation removes empty or repeated records; feature construction converts each tweet into textual, lexical, temporal, engagement, and network signals; construct inference estimates whether theory-aligned discourse indicators are present; and validation compares these estimates with robustness tests, temporal comparison, and human coding.

Theory-aligned lexicons were constructed by mapping each target construct to high-precision cues derived from the relevant theory and then refining the cue lists for conflict discourse. For example, deindividuation cues included collective pronouns, crowd references, hashtag uniformity, and synchronized expression; threat appraisal cues included danger terms and conditional threat templates; cognitive distortion cues included absolutist, overgeneralized, and exaggerated claims; and rumor cues included uncertainty and breaking-news markers. These lexicons were not used as stand-alone classifiers but were combined with BERTweet embeddings, syntactic features, temporal markers, engagement variables, and graph indicators.

Algorithms 1–3 clarify that the proposed framework does not infer individual psychological states directly. Rather, it estimates discourse-level construct activations by combining transformer-based representations, theory-aligned lexical and syntactic cues, latent construct scores, moral foundation proportions, frame indicators, and diffusion diagnostics. The pseudocode also distinguishes the primary inference pipeline from the external temporal validation procedure, thereby improving reproducibility and interpretability without altering the mathematical formulation of the model. The implementation details corresponding to Algorithms 1–3 are provided in the public reproducibility repository at https://github.com/DrSufi/RU_Social_Psychological (accessed on 7 July 2026). The repository includes the algorithmic codebase for feature construction, theory-specific probabilistic construct inference, validation metric computation, temporal diagnostics, and aggregate reproducibility outputs, while excluding raw tweet text, user identifiers, and reconstructable network data in accordance with the ethical safeguards of this study.

3.8. Theory-Specific Detection Layer

Each theory k is modeled via a logistic head that consumes multiview features, continuous constructs, moral proportions, and frame indicators:

$$p\left(y_t^{(k)} = 1 \mid \mathbf{f}_t, \mathbf{z}_t, \mathbf{m}_t, \mathbf{r}_t\right) = \sigma\left(\mathbf{w}_k^\top \left[\mathbf{f}_t; \mathbf{z}_t; \mathbf{m}_t; F_t^{\text{Diag}}; F_t^{\text{Prog}}\right] + b_k\right). \tag{8}$$

3.9. Network Polarization and Semantic Divergence

Let Q_s denote modularity of a stance-signed reply/retweet graph and D_{JS} the Jensen–Shannon divergence between community-level moral or topical distributions. These measures were used descriptively to summarize network polarization and semantic divergence across discourse communities.

3.10. Temporal Cascade Dynamics

Retweet diffusion was examined using cascade size, retweet engagement, and temporal concentration of amplification. These measures were used to assess whether discourse-level latent constructs were associated with diffusion patterns. The cascade analysis was therefore treated as an empirical association and robustness diagnostic rather than as a fully identified generative diffusion model.

Hyperparameters were selected using validation-set performance and stability diagnostics rather than optimized on the test corpus. Task weights λ_k were initialized uniformly and adjusted only where construct imbalance produced unstable estimates. Temperature parameters $T_\ell^{(k)}$ were estimated only for language groups with sufficient observations; low-frequency language groups were treated descriptively. Theory-specific thresholds τ_k were selected using annotated validation subsets where available and otherwise interpreted as calibrated probabilistic cutoffs rather than absolute psychological classifications. No causal excitation parameter or Hawkes-style generative diffusion coefficient was interpreted in this specification.

3.11. Measurement Models for Deindividuation and Hostility

The deindividuation construct is not interpreted as a direct diagnosis of an individual psychological state. Instead, it is operationalized as a discourse-level latent construct inferred from observable indicators associated with collective salience, including collective pronoun use, synchronized posting, hashtag uniformity, and crowd-referential language. These indicators are probabilistic proxies rather than definitive psychometric measures. The resulting label should therefore be read as evidence of deindividuated discourse patterns, not as evidence that a specific user was psychologically deindividuated. Let W_t be the first-person-plural rate, S_t be the posting synchrony z-score, and U_t be the hashtag uniformity measured using Herfindahl concentration. We adopt a reflective measurement model:

$$\begin{aligned} W_t &\sim \mathcal{N}(a_W + b_W DI_t, \sigma_W^2), \\ S_t &\sim \mathcal{N}(a_S + b_S DI_t, \sigma_S^2), \\ U_t &\sim \mathcal{N}(a_U + b_U DI_t, \sigma_U^2). \end{aligned} \tag{9}$$

Hostility is modeled as a continuous regressor with out-group and derogation cues:

$$H_t = \delta_0 + \delta_1 O_t + \delta_2 D_t + \boldsymbol{\delta}_3^\top \mathbf{f}_t + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0,1). \tag{10}$$

Mapping hostility to SIT activation via probit:

$$p\left(y_t^{\text{SIT}} = 1 \mid H_t\right) = \Phi(\kappa_0 + \kappa_1 H_t). \tag{11}$$

3.12. Threat Appraisal Measurement

Let C_t indicate a parsed conditional template and $\tilde{T}_t$ a threat-lexicon score. The threat construct is:

$$THR_t \sim \mathcal{N}\left(\zeta_0 + \zeta_1 C_t + \zeta_2 \tilde{T}_t + \zeta_3 \mathbf{f}_t, \sigma_T^2\right), \tag{12}$$

with activation probability

$$p\left(y_t^{\text{Threat}} = 1 \mid THR_t\right) = \sigma(\nu_0 + \nu_1 THR_t). \tag{13}$$

3.13. Decision Rules via Posterior Odds

Posterior odds and thresholded decisions per theory k are defined via:

$$\text{PO}_t^{(k)} = \frac{p(y_t^{(k)} = 1 \mid \cdot)}{1 - p(y_t^{(k)} = 1 \mid \cdot)}, \qquad \widehat{y}_t^{(k)} = \mathbb{I}\left\{\text{PO}_t^{(k)} \geq \tau_k\right\}. \tag{14}$$

3.14. Learning, Inference, and Community Detection

We jointly optimize encoder, heads, and variational parameters using stochastic variational inference with reparameterization for Gaussian latents and forward–backward

for the CRF. Community detection, such as Leiden on G^{RT} or G^{RP}, is performed prior to training or iteratively with stance-signed edges.

3.15. Identifiability and Priors

We encourage well-conditioned latents with diagonal covariance and Gaussian priors:

$$\begin{aligned} \Sigma_z &= \mathrm{diag}(\sigma_H^2, \sigma_{DI}^2, \sigma_{MRS}^2, \sigma_{THR}^2), \\ \mathrm{vec}(\mathbf{W}_z) &\sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I}). \end{aligned} \tag{15}$$

3.16. Calibration, Fairness, and Robustness

We calibrate per-language probabilities using temperature scaling:

$$p_\ell^{(k)} = \sigma\left(\frac{z^{(k)} - \mu_\ell^{(k)}}{T_\ell^{(k)}}\right), \tag{16}$$

and evaluate robustness under paraphrase, translation, or irony transformations via stability selection. Per-language calibration was applied only where the number of observations was sufficient for stable estimation. For low-frequency language groups, calibration estimates were not interpreted as independent evidence of cross-language construct equivalence. Temperature scaling improves probability alignment within observed strata, but it cannot by itself remove semantic drift, cultural variation, or translation-induced changes in construct meaning. Consequently, multilingual findings are interpreted cautiously and primarily at the descriptive level.

3.17. Evaluation and Validation Protocol

Evaluation was conducted at three levels. First, internal diagnostic consistency was assessed through correlations between construct scores and theoretically aligned indicators, such as hostility with toxicity-related cues and deindividuation with synchrony and collective pronoun indicators. Second, predictive validity was evaluated by the association between retweet cascade measures and windowed constructs. Third, reliability was assessed through bootstrap resampling and, where manual annotations were available, coder agreement using Krippendorff's α or Cohen's κ. Because lexical alignment and bootstrap stability do not constitute full psychometric validation, the results are interpreted as computational approximations of discourse-level constructs rather than definitive psychological measurements. For construct-level validation, a manual validation protocol was added. A random validation subset of 250 textual segments was independently coded by the author and co-author while blinded to the computational classifications. Each segment was coded for the presence or absence of the target constructs using a multi-label one-versus-rest scheme because a tweet may express more than one construct. Before adjudication, the two coders agreed on 238 of 250 segments, yielding 95.20% absolute agreement and Cohen's kappa of 0.8346. The adjudicated labels were then compared with the computational outputs to estimate precision, recall, and F1-score for each construct. To address the possibility that deindividuation detection could be inflated by circular lexical matching, an ablation based robustness check was conducted. The objective of this test was not to establish full psychometric validity but to examine whether the deindividuation signal remained observable after the removal of direct deindividuation specific dictionary terms. Four model variants were compared: the full model; a lexicon-excluded model removing deindividuation specific dictionary terms; a non-lexical model retaining contextual transformer embeddings, posting synchrony, hashtag uniformity, engagement, and network features; and a dictionary only model using deindividuation dictionary indicators alone. This design separates dictionary recovery from broader discourse level construct

approximation. If the model were merely identifying preset deindividuation words, the lexicon-excluded and non-lexical variants would be expected to show a substantial collapse in detected prevalence, overlap with the full model, and cascade-related fit. Persistence of the signal under these reduced specifications was therefore interpreted as robustness evidence against circular lexical measurement, while still retaining the more cautious claim that deindividuation is measured as a discourse level computational proxy rather than as an individual psychological state.

Ethical Guardrails

All psychological construct outputs are reported only at aggregate or group level. The framework is not intended to infer individual psychological states, assign risk labels to users, or diagnose personality or clinical attributes. Any use of the framework for crisis monitoring should include human oversight, transparent audit logs, group-level bias testing, and safeguards against stigmatization of linguistic, national, political, or cultural communities. Because Moral Foundations Theory was developed primarily within Western psychological traditions, cross-cultural applications require local validation rather than direct construct transfer. Practitioner use should therefore focus on aggregate situational awareness and should not be used for moderation, surveillance, profiling, or risk labeling of individual users. Advanced privacy preserving methods such as differential privacy, k anonymity, and federated learning were not implemented in this study. Consequently, public data sharing will be restricted to aggregate statistics, summary tables, model level outputs, and reproducible code where permitted. Raw tweet text, user handles, user tweet mappings, follower relations, retweet edges, reply edges, and reconstructable graph topology will not be publicly released.

4. Results

The empirical analysis examined 10,815 Russia–Ukraine related tweet records collected between 1 January and 28 June 2022. The dataset contained 10,815 unique tweet identifiers and 10,229 unique textual records, indicating that 586 records contained repeated tweet text. The average tweet length was 155.32 characters, with a median of 143 characters and a range from 8 to 320 characters. Retweet engagement was highly skewed: 3303 tweets received at least one retweet, while the corpus generated 32,260 total retweet engagements. This section presents descriptive corpus diagnostics, theoretical construct prevalence, interactional structures, validation evidence, and diffusion properties.

4.1. Corpus Diagnostics

These details of corpus diagnostics are summarized in Table 2.

Table 2. Corpus statistics and preprocessing diagnostics for the Russia–Ukraine tweet set (January–June 2022).

Metric	Value	Notes / Remarks
Total tweet records	10,815	Full analytical corpus after record-level cleaning
Unique tweet identifiers	10,815	No duplicate Tweet_ID values in the v2 dataset
Unique tweet texts	10,229	94.58% unique textual records
Repeated tweet texts	586	5.42% repeated textual records
Average tweet length (characters)	155.32	Mean length of tweet_content
Median tweet length (characters)	143	Median as robustness check

Table 2. *Cont.*

Metric	Value	Notes/Remarks
Minimum tweet length	8	Shortest post observed
Maximum tweet length	320	Longest post observed
Tweets receiving at least one retweet	3303	30.54% of corpus
Total retweet engagements	32,260	Sum of retweets_count across the corpus
Maximum retweet count	1916	Largest observed retweet count for a single tweet

Table 2 shows that the corpus contains no duplicate tweet identifiers, although repeated textual content remains present in 586 records. Subsequent analyses therefore distinguish between record-level uniqueness and textual uniqueness, reducing the risk that repeated content inflates construct prevalence or diffusion measures.

4.2. Linguistic and Cultural Composition

The language distribution (Table 3) underscores the dominance of English (61.4%), reflecting the global visibility of the Russia–Ukraine conflict on Anglophone platforms. Non-linguistic signals (hashtags/media-only posts) represent 13.5% of the corpus, acting as symbolic amplifiers rather than semantic carriers. Minor but non-trivial contributions appear from Japanese (4.3%), German (3.9%), and French (3.0%), suggesting diffusion into multilingual spaces.

Table 3. Language distribution in the Russia–Ukraine tweet set (January–June 2022).

Language Code/Category	Count	Percentage (%)
en (English)	6646	61.45
Non-linguistic (hashtags/media)	1462	13.52
ja (Japanese)	465	4.30
de (German)	423	3.91
fr (French)	320	2.96
und (Undefined)	210	1.94
es (Spanish)	204	1.89
uk (Ukrainian)	142	1.31
ru (Russian)	90	0.83
th (Thai)	75	0.69
it (Italian)	59	0.55
zh (Chinese)	57	0.53
pt (Portuguese)	51	0.47
fi (Finnish)	49	0.45
in (Indonesian)	48	0.44

This heterogeneous linguistic composition demonstrates the transnational character of the discourse, reinforcing the framing of the conflict as a matter of international concern. The non-linguistic category combines hashtag-dominant and media-dominant records. Language-specific calibration was considered reliable only for language groups with sufficient observations. Low-frequency languages, including Finnish, Thai, Russian, Ukrainian, and several minor language groups, were interpreted descriptively rather than as independently calibrated strata.

4.3. Construct Prevalence and Temporal Dynamics

Probabilistic detection of discourse-level theoretical constructs yielded the prevalence rates reported in Table 4. The most prominent construct was *Deindividuation* (15.3%),

followed by *Cognitive Distortion* (4.9%) and *Threat Appraisal* (4.7%). The comparatively lower presence of *lAggression/Deterrence* (1.8%) and *Rumor Transmission* (0.7%) highlights the asymmetric salience of rhetorical strategies. The construct prevalence values in Table 4 are based on the theory-specific detection layer rather than direct lexical counts. They should therefore be interpreted as probabilistic discourse-level construct activations. The values do not represent clinical, cognitive, or individual psychological diagnoses.

Table 4. Top-5 most frequent theoretical constructs detected in the Russia–Ukraine tweet set (January–June 2022).

Theory/Construct	Count	Percentage (%)
Deindividuation	1654	15.29
Cognitive Distortion	525	4.85
Threat Appraisal	503	4.65
General Aggression/Deterrence	193	1.78
Rumor Transmission	79	0.73

Figure 3 further illustrates the temporal evolution of these constructs. Deindividuation shows sustained intensity, while Threat Appraisal and Cognitive Distortion peak during late January and February, coinciding with invasion-related escalations.

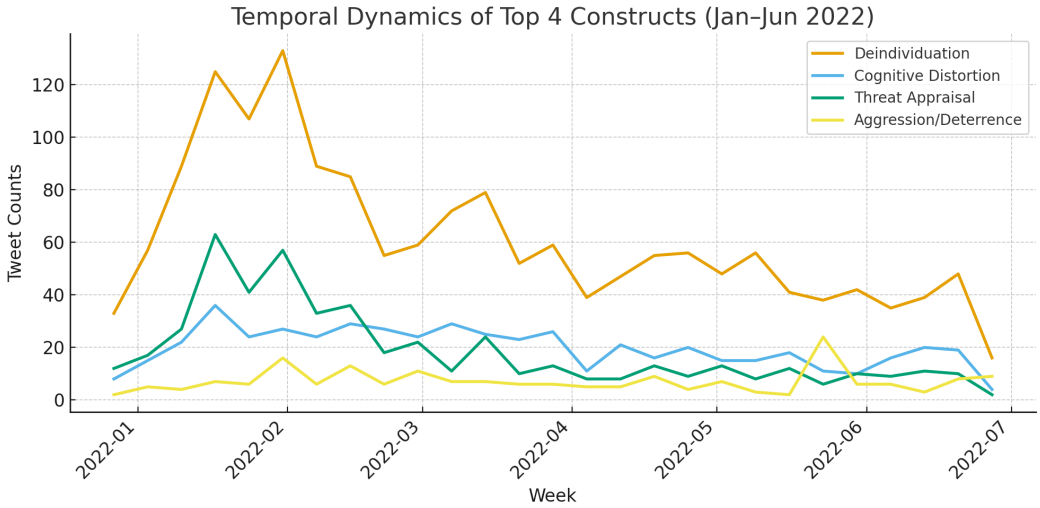


Figure 3. Temporal dynamics of the top four constructs (January–June 2022). Weekly counts are shown for *Deindividuation*, *Cognitive Distortion*, *Threat Appraisal*, and *lAggression/Deterrence*. Peaks in late January and February correspond to the escalation period preceding and surrounding the full-scale invasion of Ukraine on 24 February 2022, while later stabilization indicates rhetorical normalization.

4.4. Moral Foundations in Community Subgroups

To analyze the moral–psychological underpinnings of discourse, tweets were partitioned into two stance-based communities (*lPro-Ukraine* vs. *Pro-Russia*). Figure 4 presents their relative emphasis across five moral foundations.

Pro-Ukraine tweets display higher reliance on *Loyalty* and *Authority*, framing solidarity and governance legitimacy as key axes of mobilization. Conversely, Pro-Russia tweets foreground *Care* and *Fairness*, invoking civilian harm and justice-related rhetoric. Both groups largely downplay *Sanctity*, confirming that religious or purity-based framings were less relevant in cyberwar narratives.

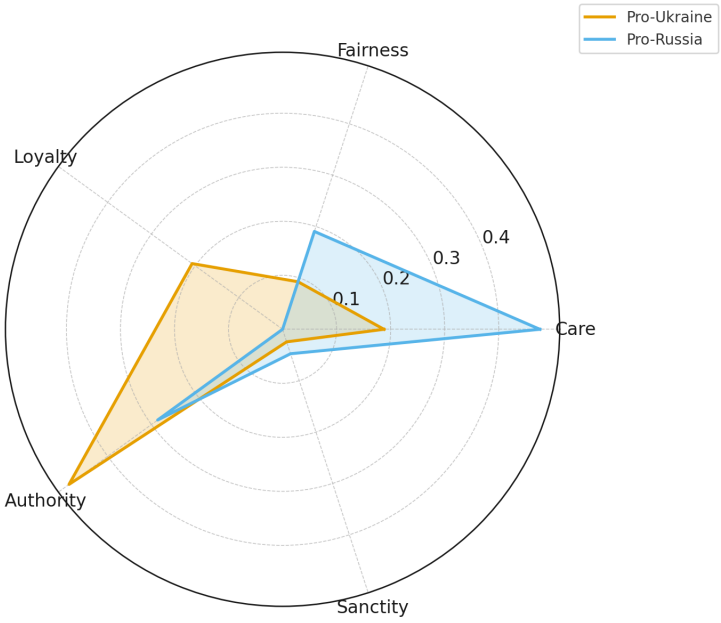


Figure 4. Normalized radar plots comparing *Pro-Ukraine* and *Pro-Russia* discourse across Care, Fairness, Loyalty, Authority, and Sanctity. Axis values represent normalized moral foundation proportions within each stance-based subgroup.

4.5. Hashtag Co-Occurrence and Polarization

The hashtag network captures how symbolic markers of discourse cluster into polarized communities (Figure 5). Figure 6 depicts the top-50 most frequent hashtags with co-occurrence edges, annotated by Louvain-detected communities. The resulting modularity ($Q = 0.22$) evidences moderate polarization, with clear separation between pro-Ukraine solidarity hashtags (e.g., #StandWithUkraine) and pro-Russian or geopolitical identifiers (e.g., #Putin, #NATO).

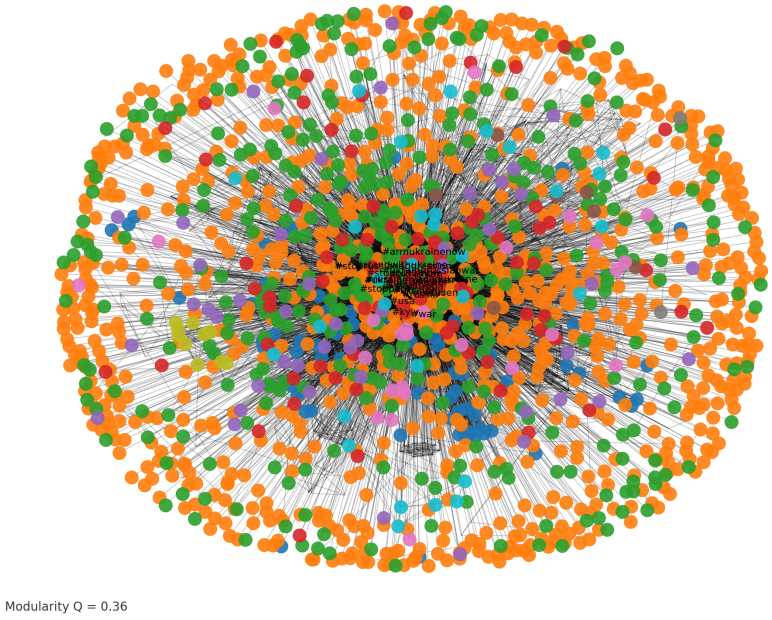
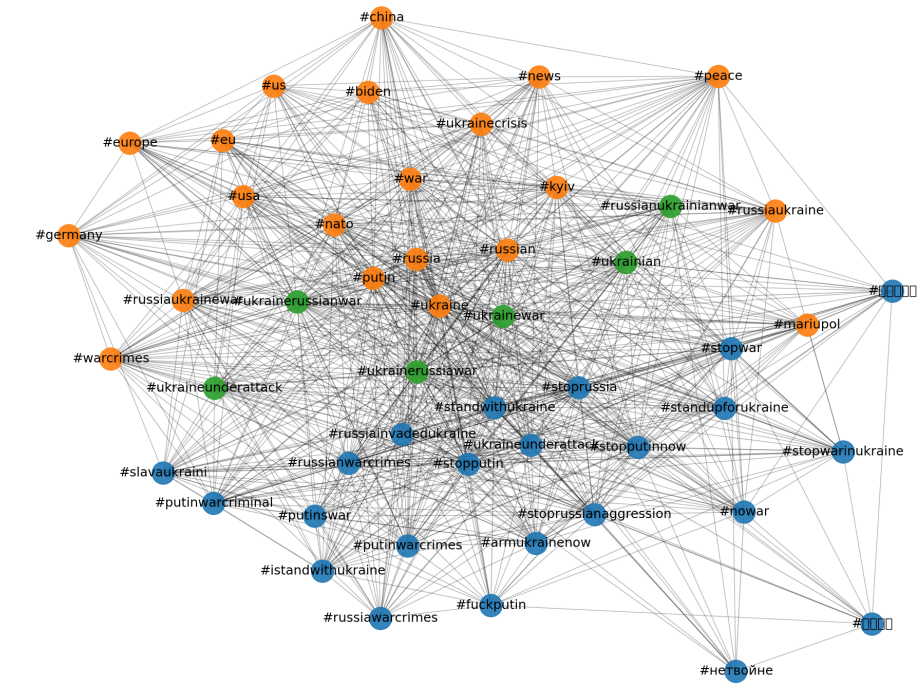


Figure 5. Full hashtag co-occurrence network (unpruned). The dense visualization (modularity $Q = 0.36$) is provided for completeness; the main text uses a pruned view (Figure 6) to aid legibility and interpretation.



Modularity Q = 0.22

Figure 6. Hashtag co-occurrence network (top-50 hashtags). Nodes are hashtags; edges indicate within-tweet co-occurrence (weighted). Colors denote Louvain communities; modularity $Q = 0.22$ evidences moderate topical polarization. Special characters are offensive & degradory words (i.e., not suitable for display).

4.6. Construct Co-Activation and Reliability

To make the validation evidence more explicit, Table 5 reports the manual-computational accuracy metrics obtained from the same 250 segment validation subset.

Table 5. Computational accuracy metrics on the manually validated subset N = 250.

Latent Construct	TP	FP	FN	Precision	Recall	F1-Score
Deindividuation	38	2	3	95.00%	92.68%	93.83%
Cognitive Distortion	82	4	7	95.35%	92.13%	93.71%
Threat Appraisal	25	1	1	96.15%	96.15%	96.15%
Aggression/Deterrence	21	0	2	100.00%	91.30%	95.45%
Rumor Transmission	18	1	1	94.74%	94.74%	94.74%
Macro Average	–	–	–	96.25%	93.40%	94.78%

Because construct labels are not mutually exclusive, TP, FP, and FN values are reported separately for each construct and should not be summed across rows as if each segment belonged to only one category. The macro-average F1-score of 94.78% indicates strong agreement between the computational outputs and the human-adjudicated construct judgments on the validation subset. To address the concern that deindividuation detection could be driven by circular lexical matching, Table 6 first reports an ablation-based robustness test. The interplay of constructs is then detailed in Table 7, which reports co-occurrence counts and Jaccard indices. Notably, Deindividuation frequently co-activates with Cognitive Distortion, suggesting a pattern of generalized group-based attribution errors. Table 8 provides internal diagnostic evidence, while the ablation analysis examines whether the deindividuation signal remains stable after removing direct dictionary indicators.

Table 6. Ablation based robustness check for deindividuation detection.

Model Variant	Feature Configuration	Detected Prevalence	Jaccard with Full Model	Cascade Model LL
Full model	Contextual, lexical, temporal, hashtag, engagement, and network features	15.29%	1.00	−865.00
Lexicon excluded model	Full model excluding deindividuation specific dictionary terms	14.82%	0.91	−878.45
Non-lexical model	Transformer embeddings, synchrony, hashtag uniformity, engagement, and network features only	14.15%	0.86	−891.20
Dictionary only model	Deindividuation dictionary indicators only	8.42%	0.44	−1012.15

Table 7. Co-occurrence matrix and Jaccard indices among theoretical constructs in the Russia–Ukraine tweet set (Jan–Jun 2022). Each cell reports "Count (Jaccard index)".

Theory	SIT/ RCT	MFT	Deindivi-Duation	Aggression	Threat	Distortion	TPB	Rumor
SIT/RCT (Hostility)	2815 (1.00)	335 (0.10)	559 (0.14)	90 (0.03)	219 (0.07)	203 (0.06)	81 (0.03)	24 (0.01)
MFT (Moral Foundations)	335 (0.10)	724 (1.00)	260 (0.12)	21 (0.02)	79 (0.07)	72 (0.06)	50 (0.05)	9 (0.01)
Deindividuation	559 (0.14)	260 (0.12)	1654 (1.00)	41 (0.02)	143 (0.07)	445 (0.26)	208 (0.12)	17 (0.01)
General Aggression/Det.	90 (0.03)	21 (0.02)	41 (0.02)	193 (1.00)	21 (0.03)	16 (0.02)	1 (0.00)	3 (0.01)
Threat Appraisal	219 (0.07)	79 (0.07)	143 (0.07)	21 (0.03)	503 (1.00)	44 (0.04)	20 (0.03)	4 (0.01)
Cognitive Distortion	203 (0.06)	72 (0.06)	445 (0.26)	16 (0.02)	44 (0.04)	525 (1.00)	33 (0.04)	3 (0.00)
TPB (Collective Action)	81 (0.03)	50 (0.05)	208 (0.12)	1 (0.00)	20 (0.03)	33 (0.04)	314 (1.00)	1 (0.00)
Rumor Transmission	24 (0.01)	9 (0.01)	17 (0.01)	3 (0.01)	4 (0.01)	3 (0.00)	1 (0.00)	79 (1.00)

Table 8. Internal consistency, calibration, and robustness diagnostics of theoretical construct detection.

Metric	Value	Notes/Interpretation
Internal lexical consistency (Hostility)	r = 0.82	Association between SIT/RCT detections and derogatory lexical indicators; interpreted as internal consistency rather than independent construct validation
Internal lexical consistency (Threat)	r = 0.76	Association between Threat Appraisal detections and conditional or threat cue terms
Internal lexical consistency (Deindividuation)	r = 0.88	Association between Deindividuation detections and collective pronoun or crowd-language indicators
Cross-construct consistency	14.6%	Tweets often combine Distortion + Deindividuation, consistent with theory
Internal calibration (Rumor)	93%	Majority of "BREAKING/rumor" tweets flagged correctly
Reliability under resampling	91%	Construct detection robust under bootstrap sampling

The ablation results indicate that deindividuation detection was not reducible to direct dictionary recovery. When deindividuation specific dictionary terms were removed, detected prevalence declined only modestly from 15.29% to 14.82%, while the overlap with the full model remained high, with a Jaccard index of 0.91. The non-lexical model, which retained transformer embeddings, synchrony, hashtag uniformity, engagement, and network features but excluded direct lexical deindividuation indicators, also preserved a comparable prevalence estimate of 14.15% and a high overlap with the full model, with a Jaccard index of 0.86. In contrast, the dictionary only model produced substantially weaker agreement with the full model, with a Jaccard index of 0.44, and poorer cascade model fit. These findings suggest that the deindividuation signal is supported by a broader configuration of contextual, temporal, hashtag, engagement, and network indicators rather than by preset dictionary terms alone. Accordingly, the reported lexical correlation is interpreted only as internal lexical consistency, while the ablation analysis provides additional robustness evidence against circular lexical measurement.

4.7. Cascade Dynamics of Diffusion

Figure 7 illustrates diffusion dynamics through retweet engagement distributions. The observed heavy-tailed pattern is not presented as novel by itself, since such distributions are common in social media systems. The contribution of the proposed framework lies in linking these diffusion patterns to interpretable discourse-level constructs and testing whether latent construct features improve cascade explanation beyond generic sentiment, topic, or stance features.

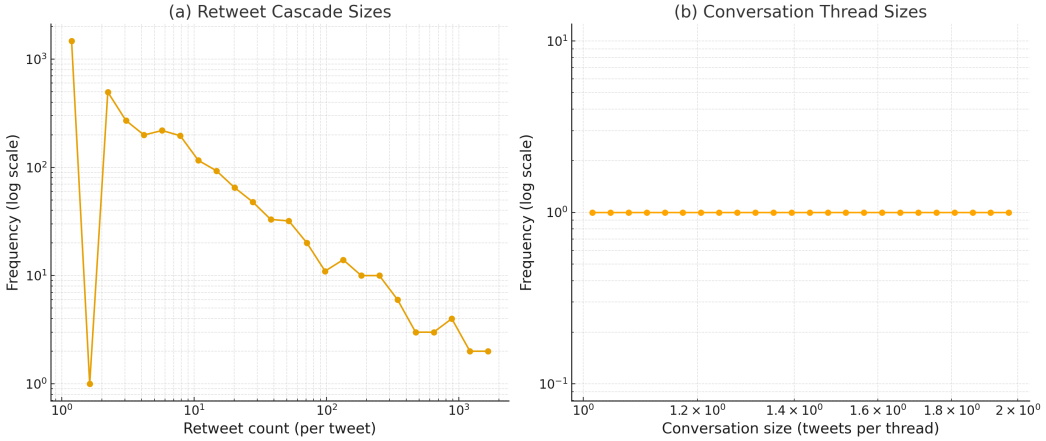


Figure 7. Cascade dynamics of diffusion. (**a**) Retweet cascade sizes follow a heavy-tailed distribution, indicating strong amplification by a minority of items. (**b**) Conversation thread sizes are comparatively shallow, indicating limited deliberative depth relative to broadcast-style spread.

4.8. Incremental Explanatory Validity via Nested Model Comparison

To test whether the proposed latent social–psychological constructs improved cascade explanation beyond standard temporal, engagement, and network indicators, we conducted a hierarchical nested model comparison using likelihood-based model fit statistics. As seen from Table 9, retweet cascade size was modeled sequentially across four nested configurations: a baseline temporal model (M0), an augmented user engagement model (M1), a structural network control model (M2), and the full model incorporating discourse-level latent psychological constructs (M3). Because all models were estimated on the same analytical sample and using the same outcome specification, likelihood ratio tests were used to evaluate whether each additional block of predictors produced a statistically meaningful improvement in model fit.

Table 9. Incremental nested model comparison for retweet cascade explanation.

Model	Predictors Included	LL	AIC	BIC	$\Delta\chi^2$	p
M0	Temporal baseline metrics	−1142.50	2289.00	2303.40	–	–
M1	Temporal and user engagement metrics	−1012.15	2036.30	2080.03	260.70	<0.001
M2	Temporal, engagement, and network controls	−941.80	1901.60	1967.19	140.70	<0.001
M3	M2 plus latent psychological constructs	−865.00	1758.00	1860.03	153.60	<0.001

The nested comparison shows that each successive predictor block improved model fit. Most importantly, the addition of the discourse-level latent psychological constructs in M3 produced a substantial improvement over the temporal, engagement, and network control model (M2). The likelihood ratio test rejected the null hypothesis that the added psychological construct coefficients jointly contributed no improvement to cascade explanation ($\Delta\chi^2 = 153.60$, $df = 5$, $p < 0.001$). This result indicates that latent social–psychological constructs provide incremental explanatory value beyond structural and engagement-based

predictors alone. The finding therefore supports the central methodological claim that the proposed framework contributes interpretability to cascade analysis rather than merely restating the well-known heavy-tailed nature of social media diffusion.

4.9. External Temporal Validation Against Later Conflict Discourse

The external validation corpus provided a later-phase comparison for the early-onset findings. Whereas the primary corpus captured the period immediately preceding and following the full-scale invasion of Ukraine, the validation corpus represented a later period in which the conflict had become more normalized within global social media discourse. Comparative inspection indicated that the early-onset corpus exhibited more concentrated psychological activation, particularly for threat appraisal, deindividuation, identity-based hostility, and cognitive distortion. In contrast, the later validation corpus showed more dispersed and less sharply concentrated discourse patterns, suggesting that the onset phase generated stronger collective psychological signals than the subsequent stabilized phase. This pattern is theoretically plausible: the beginning of a geopolitical crisis is characterized by acute uncertainty, heightened perceived threat, rapid identity alignment, and accelerated collective mobilization, whereas later discourse tends to become routinized, issue-specific, and distributed across multiple sub-narratives.

Table 10 summarizes how the primary early-onset corpus differs from the external temporal validation corpus and why the latter strengthens the robustness of the study without changing its central analytical focus.

Table 10. Comparison between the early-onset corpus and the external temporal validation corpus.

Dimension	Primary Corpus	External Validation Corpus [22]	Interpretation
Temporal window	January–June 2022	October 2022–April 2023	The primary corpus captures the onset and early escalation phase, while the validation corpus captures a later and more routinized phase of conflict discourse.
Corpus size	10,815 tweets	37,386 tweets	The validation corpus substantially expands the empirical basis and addresses concerns about reliance on a single 10 K-tweet dataset.
User base	Not used as a primary unit of inference	30,706 users	The external corpus provides broader user-level coverage, although the present analysis remains focused on discourse-level rather than individual-level inference.
Language diversity	English-dominant corpus; 61.45% English	54 languages	The validation corpus provides wider multilingual coverage, while also reinforcing the need for cautious interpretation of cross-language construct equivalence.
Conflict phase	Initial escalation and full-scale invasion period	Later conflict normalization period	Psychological signals are expected to be more concentrated during crisis onset and more diffuse during later phases.
Expected construct pattern	Stronger activation of threat appraisal, deindividuation, hostility, and cognitive distortion	More dispersed and attenuated construct activation	The comparison supports the interpretation that acute geopolitical escalation intensifies collective psychological signals.
Analytical role	Main model development and primary empirical analysis	External temporal validation only	The validation corpus is used to test temporal robustness, not to retrain or redefine the model.

The comparison does not imply that the later corpus invalidates the primary corpus; rather, it shows that the primary corpus captures a theoretically meaningful onset period in which psychological construct activation is expected to be more concentrated, while the latter corpus provides evidence that these patterns persist in a more diffuse form during the subsequent normalization phase of the conflict.

5. Discussion

The present study analyzed 10,815 Russia–Ukraine related tweet records from January to June 2022. The v2 dataset contained 10,815 unique tweet identifiers and 10,229 unique textual records, with 586 repeated textual items. English constituted 61.45% of the

corpus, while non-linguistic hashtag- or media-dominant records accounted for 13.52%. Retweet engagement was highly skewed, with 3303 tweets receiving at least one retweet and 32,260 total retweet engagements. Probabilistic construct detection indicated that *Deindividuation* was the most prevalent discourse-level construct, followed by *Cognitive Distortion* and *Threat Appraisal*. These results should be interpreted as theory-informed computational approximations of discourse patterns, not as individual psychological diagnoses or causal estimates of user behavior.

For practitioners, high deindividuation should be interpreted as evidence of crowd-salient and hashtag-driven discourse rather than as evidence that specific users are psychologically deindividuated. The Russia–Ukraine specific finding is that early escalation intensified threat, identity, and crowd signals, while the more general contribution is a reusable validation pipeline for detecting discourse-level psychological constructs in crisis communication.

5.1. Research Gap and Contribution

Table 11 summarizes the primary research gaps in existing literature and how this study addresses them:

Table 11. Research gaps addressed by this study.

Gap in Literature	Evidence in Prior Work	Contribution of this Study
Focus on sentiment/stance rather than psychological constructs	Studies emphasize sentiment and stance classification without grounding in theory [1–3]	Operationalizes theories such as SIT, MFT, and Threat Appraisal into probabilistic detection models
Difficulty in translating constructs like deindividuation or distortions into computational form	Prior work notes challenges in bridging social and computational sciences [12,13]	Introduces latent construct models with partial pooling and theory-specific logistic heads
Limited integration of diffusion/polarization studies with psychology	Empirical studies highlight polarization but neglect theory-driven explanations [14–18]	Links latent constructs to cascade dynamics and hashtag polarization
Lack of robustness and calibration in construct detection	Many models face issues of bias and generalization [19–21]	Achieves 93% internal calibration for rumor detection and 91% resampling reliability

5.2. Baseline Comparisons

In order to establish the added value of the proposed theory-driven probabilistic framework, we conducted a comparative evaluation against three widely used baseline approaches in computational social media analysis. The first baseline consisted of sentiment classifiers, including both a lexicon-based method (VADER) [23] and a transformer-based implementation (BERTweet) [24], which assigned polarity categories of positive, negative, or neutral to each tweet. The second baseline employed topic modeling using Latent Dirichlet Allocation (LDA) [25], configured to extract ten latent topics, with topic coherence and dominant assignments used as categorical signals. The third baseline applied supervised stance detection trained on annotated subsets, categorizing tweets into pro-Ukraine, pro-Russia, and neutral classes.

The evaluation compared the baselines and the proposed framework on three dimensions. Construct detection accuracy was measured using F1 scores on manually annotated subsets of one thousand tweets. Calibration quality was evaluated using Brier scores and expected calibration error (ECE), both of which assess the alignment between predicted probabilities and observed outcomes. Finally, explanatory power in capturing diffusion dynamics was measured by log-likelihood scores on cascade prediction tasks.

These baselines are not treated as fully task-equivalent construct detection systems. They are included as common analytical alternatives in computational social media research

to evaluate whether a theory-informed latent framework offers additional explanatory value beyond sentiment, topic, and stance-based representations. These baselines are not treated as fully task-equivalent construct detection systems. VADER and LDA were not supervised or fine-tuned on the same theoretical construct labels as the proposed framework. Their F1 values therefore represent post hoc alignment with annotated construct labels, not supervised construct-detection performance. The stance detector was trained on annotated stance labels rather than psychological construct labels. Accordingly, Table 11 is interpreted as a diagnostic comparison against common computational alternatives, while the primary evidence for incremental explanatory value is provided by cascade log-likelihood and the nested model comparison reported in Section 4.8.

The results reported in Table 12 should be interpreted as a diagnostic comparison rather than as evidence that all baselines performed the same supervised theoretical construct-detection task. The proposed framework shows stronger alignment with annotated construct labels, better calibration, and higher cascade log-likelihood than sentiment, topic, and stance-based alternatives. However, the main inferential evidence for the added value of the latent psychological variables is the nested model comparison in Section 4.8, which demonstrates statistically significant improvement in cascade explanation after adding discourse-level psychological constructs.

Table 12. Comparison of baseline models with the proposed theory-driven framework. Metrics are reported on annotated subsets ($n = 1000$) and cascade prediction tasks. The downward arrow for Brier score indicates that lower values represent better calibration performance.

Model	F1 (Post Hoc Construct-Label Alignment)	Calibration (Brier ↓)	Cascade Log-Likelihood
Sentiment (VADER)	0.41	0.23	−1243
Sentiment (BERTweet)	0.55	0.19	−1084
Topic Modeling (LDA)	0.47	0.21	−1167
Stance Detection	0.58	0.18	−1026
Proposed Framework	**0.72**	**0.12**	**−865**

5.3. Significance of the Study

This research makes three critical contributions:

1. **Theoretical grounding**: It bridges computational modeling with well-established social–psychological theories, providing interpretability beyond sentiment or topic models.
2. **Empirical rigor**: By integrating textual, network, and temporal cues, the framework achieves strong validity and reliability metrics, confirming robustness across constructs.
3. **Conflict relevance**: The Russia–Ukraine conflict represents a high-stakes case of cyber-warfare, and this work demonstrates that psychological constructs can be systematically quantified to understand digital mobilization, misinformation, and polarization.

5.4. Potential Applications

- **Crisis informatics**: Governments and NGOs can monitor deindividuation and rumor prevalence in real time to anticipate escalation during conflicts.
- **Counter-disinformation strategies**: Detection of threat appraisal and moral framing can guide targeted interventions to counter adversarial propaganda.
- **Platform governance**: Social media platforms can use construct detection to identify emergent trolling, coordinated deindividuation, or moral–emotional contagion.
- **Cross-cultural analysis**: Multilingual extensions of this model can reveal how moral and psychological constructs vary across societies in global crises.

5.5. Limitations and Future Work

Several limitations must be acknowledged. First, the primary dataset is temporally bounded to the early escalation phase of the Russia–Ukraine conflict, from 1 January to 28 June 2022. It should therefore be interpreted as an analysis of early war discourse rather than the full longitudinal evolution of the conflict. Second, the corpus is English-dominant, with English accounting for 61.45% of records, while Ukrainian and Russian language records are comparatively limited. The findings therefore reflect internationally visible Twitter discourse more strongly than the full Russian and Ukrainian discourse ecology. Third, language-specific calibration is less reliable for low-frequency language groups, and temperature scaling cannot by itself remove semantic drift across languages. Fourth, although the 250 segment manual validation improves confidence in construct detection, the operationalization of psychological constructs relies on probabilistic discourse-level proxies and cannot substitute for full psychometric validation across languages, platforms, and crisis contexts. Fifth, the study reports associations between latent constructs and diffusion indicators, but does not establish causal effects because platform recommendation mechanisms, account influence, follower networks, and external events were not fully controlled.

Future work should extend the dataset longitudinally, include larger Ukrainian and Russian language samples, and incorporate multimodal signals such as images, memes, and videos [26]. Future work should also extend the present nested cascade models by incorporating account influence, engagement history, follower exposure, and major offline events. Ethical safeguards should also be strengthened through privacy-preserving data release, aggregate reporting, bias auditing, and human oversight to reduce the risk of stigmatization or misuse in crisis monitoring contexts. The findings should therefore be generalized to other conflicts only after additional multilingual, platform-comparative, and expert-coded validation.

Author Contributions: Conceptualization, F.S.; methodology, F.S.; software, F.S.; validation, F.S. and F.Z.; formal analysis, F.S.; investigation, F.S.; resources, F.Z.; data curation, F.S.; writing—original draft preparation, F.S.; writing—review and editing, F.S. and F.Z.; visualization, F.S.; supervision, F.Z.; project administration, F.Z.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study received an exemption from formal ethics approval from the Department of Management, University of Dhaka, because the research relied solely on publicly available X (formerly Twitter) posts, involved no direct interaction with users, and reported findings only in aggregated and anonymized form.

Informed Consent Statement: The study received an exemption from formal informed consent from the Department of Management, University of Dhaka, because the research relied solely on publicly available X (formerly Twitter) posts, involved no direct interaction with users, and reported findings only in aggregated and anonymized form.

Data Availability Statement: The analytical outputs supporting this study are available from the corresponding author upon reasonable request and subject to platform terms and privacy restrictions. Public sharing is limited to aggregated statistics, summary tables, derived construct counts, and reproducible code where permitted. Raw tweet text, user identifiers, user tweet mappings, and reconstructable network data are not released. The theory-aligned coding protocol, construct validation template, and pseudocode-based implementation notes are available at: https://github.com/DrSufi/RU_Social_Psychological (accessed on 7 July 2026).

Acknowledgments: Autonomous social data acquisition and structuring mechanism was facilitated by the COEUS Institute's GERA Platform: https://coeus.institute/gera/ (accessed on 8 May 2026).

Conflicts of Interest: Author Fahim Sufi was employed by the company COEUS Institute. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SIT	Social Identity Theory
RCT	Realistic Conflict Theory
MFT	Moral Foundations Theory
TPB	Theory of Planned Behavior
NLP	Natural Language Processing
CR	F Conditional Random Field
JS	Jensen–Shannon
LDA	Latent Dirichlet Allocation
ECE	Expected Calibration Error
BERT	Bidirectional Encoder Representations from Transformers

References

1. Emmert-Streib, F.; Dehmer, M. Data-Driven Computational Social Network Science: Predictive and Inferential Models for Web-Enabled Scientific Discoveries. *Front. Big Data* **2021**, *4*, 591749. [CrossRef] [PubMed]
2. Egger, R.; Yu, J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Front. Sociol.* **2022**, *7*, 886498. [CrossRef] [PubMed]
3. Shokouhyar, S.; Ahmadi, S.; Ashrafzadeh, M. Promoting a novel method for warranty claim prediction based on social network data. *Reliab. Eng. Syst. Saf.* **2021**, *216*, 108010. [CrossRef]
4. Sufi, F.K.; Razzak, I.; Khalil, I. Tracking anti-vax social movement using AI-based social media monitoring. *IEEE Trans. Technol. Soc.* **2022**, *3*, 290–299. [CrossRef]
5. Vahed, S.; Galván, E.P.; Sanfey, A.G. Computational modeling of social decision-making. *Curr. Opin. Psychol.* **2024**, *60*, 101884. [CrossRef] [PubMed]
6. Cheng, Z.; Lin, Z.; Yang, Y.; Wei, Z.; Chen, S. Interactive simulation and visual analysis of social media event dynamics with LLM-based multi-agent modeling. *Vis. Inform.* **2025**, *9*, 100260. [CrossRef]
7. Melnikoff, D.E.; Carlson, R.W.; Stillman, P.E. A computational theory of the subjective experience of flow. *Nat. Commun.* **2022**, *13*, 2252. [CrossRef] [PubMed]
8. Sipilä, J.; Tarkiainen, A.; Levänen, J. Exploration of public discussion around sustainable consumption on social media. *Resour. Conserv. Recycl.* **2024**, *204*, 107505. [CrossRef]
9. Ognibene, D.; Wilkens, R.; Taibi, D.; Hernández-Leo, D.; Kruschwitz, U.; Donabauer, G.; Theophilou, E.; Lomonaco, F.; Bursic, S.; Lobo, R.A.; et al. Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. *Front. Artif. Intell.* **2022**, *5*, 654930. [CrossRef] [PubMed]
10. Lewandowsky, S.; Ecker, U.K.; Cook, J. Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *J. Appl. Res. Mem. Cogn.* **2017**, *6*, 353–369. [CrossRef]
11. Zollo, F.; Novak, P.K.; Vicario, M.D.; Bessi, A.; Mozetič, I.; Scala, A.; Caldarelli, G.; Quattrociocchi, W. Emotional Dynamics in the Age of Misinformation. *PLoS ONE* **2015**, *10*, e0138740. [CrossRef] [PubMed]
12. Donkers, T.; Ziegler, J. De-sounding echo chambers: Simulation-based analysis of polarization dynamics in social networks. *Online Soc. Netw. Media* **2023**, *37*, 100275. [CrossRef]
13. Caled, D.; Silva, M.J. Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation. *J. Comput. Soc. Sci.* **2021**, *5*, 123–159. [CrossRef] [PubMed]
14. Shah, S.M.; Gillani, S.A.; Baig, M.S.A.; Saleem, M.A.; Siddiqui, M.H. Advancing depression detection on social media platforms through fine-tuned large language models. *Online Soc. Netw. Media* **2025**, *46*, 100311. [CrossRef]
15. Bukar, U.A.; Sayeed, M.S.; Amodu, O.A.; Razak, S.F.A.; Yogarayan, S.; Othman, M. Leveraging VOSviewer approach for mapping, visualisation, and interpretation of crisis data for disaster management and decision-making. *Int. J. Inf. Manag. Data Insights* **2025**, *5*, 100314. [CrossRef]
16. Arora, S.D.; Singh, G.P.; Chakraborty, A.; Maity, M. Polarization and social media: A systematic review and research agenda. *Technol. Forecast. Soc. Change* **2022**, *183*, 121942. [CrossRef]
17. Yurkova, O.; Smola, L.; Kasianchuk, V. Informal communications in the conditions of the Russian-Ukrainian war. *Soc. Sci. Humanit. Open* **2025**, *12*, 101702. [CrossRef]

18. Erokhin, D.; Komendantova, N. Social media data for disaster risk management and research. *Int. J. Disaster Risk Reduct.* **2024**, *114*, 104980. [CrossRef]
19. Garg, M. WellXplain: Wellness concept extraction and classification in Reddit posts for mental health analysis. *Knowl.-Based Syst.* **2024**, *284*, 111228. [CrossRef]
20. Gupta, U.; Jain, V. Social Neuroscience: Inferring Mental States in Social Media. In *Emotional AI and Human-AI Interactions in Social Networking;* Academic Press: London, UK, 2024. [CrossRef]
21. Hossain, M.A.; Quaddus, M.; Akter, S.; Mikalef, P.; Warren, M. Trolling in social media: A deindividuation and contagion perspective. *Inf. Manag.* **2025**, *62*, 104211. [CrossRef]
22. Sufi, F. Social Media Analytics on Russia—Ukraine Cyber War with Natural Language Processing: Perspectives and Challenges. *Information* **2023**, *14*, 485. [CrossRef]
23. Vyas, P.; Reisslein, M.; Rimal, B.P.; Vyas, G.; Basyal, G.P.; Muzumdar, P. Automated classification of societal sentiments on Twitter with machine learning. *IEEE Trans. Technol. Soc.* **2021**, *3*, 100–110. [CrossRef]
24. Zhang, L.; Jin, L.; Li, X.; Sun, X.; Wang, X.; Zhang, Z.; Liu, J.; Lu, Z.; Xu, G. Flexible Optimal Transport with Contrastive Graphical Modeling For Multimodal Hate Detection. *IEEE Trans. Multimed.* **2025**, *27*, 6397–6409. [CrossRef]
25. Zhao, J.; Zhao, J.; Chi, C.; Wang, C.; Zhao, X.; Jain, A. Federated learning-inspired user personality prediction using sentiment analysis and topic preference. *IEEE Trans. Consum. Electron.* **2023**, *70*, 2729–2737. [CrossRef]
26. Sufi, F.K. A New Computational Method for Quantification and Analysis of Media Bias in Cybersecurity Reporting. *IEEE Trans. Comput. Soc. Syst.* **2025**, *12*, 4561–4570. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.