Identity Without Action: Rethinking Collective Action Models in Disinformation Research

Lorella Viola
Vrije Universiteit Amsterdam / De Boelelaan 1105
1081 HV
Amsterdam, The Netherlands
l.viola@vu.nl

Abstract

Despite the rapid growth of disinformation research, the fundamental reasons behind user engagement with such content remain poorly understood. Recently, several scholars have suggested that researchers should study engagement with disinformation as a form of collective action (CA). Drawing on Social Identity Theory (SIT) and the Social Identity Model of Collective Action (SIMCA), this study empirically verifies this assumption by testing it across two distinct linguistic communities, English and Spanish. Specifically, it investigates whether mobilizing CA language functions as a uniform predictor of engagement, or if engagement is primarily driven by community specific identity dynamics. The experiment analysed a bilingual corpus of 4,035 X (formerly Twitter) posts associated with conspiracy theory and disinformation-related hashtags (e.g., #Agenda2030, #TheGreatReset). Using a mixed-methods approach combining BERTopic for narrative discovery, non-parametric statistical testing and Random Forest Regressor, we disentangled the effects of language presence from community behaviour. The results reveal that the Spanish community exhibits a higher baseline engagement compared to the English community indicating that engagement is primarily driven by macro-level community norms (i.e., identity) rather than micro-level linguistic triggers. We argue that rather than treating mobilizing language as a uniform predictor of engagement, future application of SIMCA in disinformation research should account for these identity-based baseline differences.

1 Introduction

Online disinformation, understood here as the deliberate dissemination of false or misleading content with the potential to cause public harm (Tucker et al., 2018), has surged in recent years, particularly in the wake of the COVID-19 crisis. Widely acknowledged as a major threat to public and individual safety, substantial research has examined its structure (Van Prooijen and Douglas, 2018; Van Prooijen and Van Vugt, 2018), spread (Bonnevie et al., 2021), impact (Simms et al., 2020; Stabile et al., 2019; Chen et al., 2020), and content (Wiggins, 2023; Demata et al., 2022; Fallis, 2009). Scholars have also investigated the role of conspiracy theories and disinformation in shaping public perceptions and decision-making (Chen et al., 2021; Yagi et al., 2024), as well as the influence of network structures and user interactions in amplifying disinformation (Quintana et al., 2022; Gunaratne et al., 2019). More recently, attention has turned to understanding the deeper mechanisms through which disinformation persuades individuals to accept unlikely or false narratives, giving more importance to the cognitive and identity-based factors that may explain why individuals engage with disinformation (Reddi et al., 2023; Bastick, 2021; Butter and Knight, 2020).

Despite the rapid growth of this field, however, the fundamental reasons behind user engagement with disinformation remain poorly understood. Some researchers argue that this gap stems from contradictory and fragmented findings (Birchall and Knight, 2022; Kirchner and Reuter, 2020), while others point to the overly functionalist approach to disinformation, treating it as merely 'the opposite of true' and overlooking its cognitive and subjective dimensions (Viola, 2025b; Reddi et al., 2023; Bastick, 2021). This often leads to disinformation consumers being dismissed as irrational or paranoid actors and to counter-measures being mostly inefficient (Alava et al., 2017; Conway, 2017; Johnson, 2018; Mølmen and Ravndal, 2021; Reicher and Haslam, 2016). This scholarship also contends that engagement with disinformation is not merely an act of passive belief but an active discursive process, where individuals construct and negotiate their cultural identities. It further argues that interacting with disinformation, through

Proceedings of SIGHUM 2026, pages 36–47

shares, comments, reactions, and reposts, constitutes a discursive practice that equally encodes collective agency. This behaviour would be shaped by perceived injustice, disillusionment with mainstream media, and the affordances of alternative information ecosystems (Wintterlin et al., 2023).

These arguments are supported by two key considerations. First, much of the current research on disinformation remains disproportionately focused on a small subset of industrialized democracies, particularly the United States and the United Kingdom (Bajaj, 2024). A study by Seo & Faris (Seo and Faris, 2021) found that 62.8% of empirical studies published in communication journals between 2015 and 2020 relied on U.S.-based data (p. 1166). Scholars such as Bajaj (Bajaj, 2024) highlight that disinformation is not a universal phenomenon and that this geographical bias distorts our understanding of its cultural dimensions. Consequently, mitigation efforts that ignore these cultural dynamics risk being ineffective.

Second, disinformation has been linked to citizens' decrease of trust in mainstream media and other sources of authoritative information (MacFarquhar, 2016; Lewis and Marwick, 2017; Allcott and Gentzkow, 2017), mistrust of establishment political figures and institutions and increased acceptance of, or indeed support for, fringe, anti-establishment or radical actors and movements (Beauchamp, 2019; Amlinger, 2022; Reichardt, 2022). Research on the 2020 health crisis, for example, has demonstrated that COVID-19 disinformation motivated individuals to protest by offering a sense of agency and empowerment (Reichardt, 2022; Amlinger, 2022; Birchall and Knight, 2022). Thus, according to this view, even unlikely or improbable disinformation narratives succeed in mobilizing individuals by fostering a belief in their capacity to effect change. In this sense, online participation, e.g., expressing dissent, signing petitions, or sharing content, would function as a low-cost form of collective action (CA) (Brunsting and Postmes, 2002). Through social media, users would be able to signal group membership, express opposition to elites, and reinforce a shared identity, transforming engagement with disinformation into a performative act of resistance. Online activities such as sharing or liking would in this way offer an easy and effective way for people to express dissent with others or to demonstrate their belonging to a group.

Building on this literature, the present study investigates engagement with disinformation on social media through the lens of social identity and CA, while explicitly questioning whether these frameworks adequately capture platform-mediated interaction. Rather than assuming that observable engagement reflects intentional mobilization or participatory efficacy, the study examines whether CA language functions as a universal engagement booster, or if its apparent effect is a byproduct of the higher baseline activity inherent to specific linguistic communities. To this end, we analyse a bilingual corpus of 4,085 English and Spanish posts from X associated with disinformation and conspiracy theory narratives. By combining BERTopic for narrative discovery with non-parametric statistical testing (Mann-Whitney U) and Random Forest Regressor to predict engagement levels, we assesses whether call-to-action CTA) vocabulary provides a consistent engagement effect (i.e., likes, reposts, quotes, replies) across groups, or if engagement is primarily driven by identity-based community behaviour. This distinction is central in refining how CA models are applied to social networks and reassessing the validity of engagement metrics as proxies for real-world mobilization connected to disinformation.

2 User engagement, disinformation, and collective action

The scholarly literature on user engagement in online and social media contexts has approached this phenomenon through various conceptual frameworks and methodologies. Engagement is often conceptualized as user-initiated actions that contribute to value co-creation, as proposed by Brodie et al. (Brodie et al., 2013). This broad definition underscores the interplay between behavioural, cognitive, and emotional dimensions of engagement, emphasizing the need to explore its motivations and nuances. Shao (Shao, 2009) categorized user interaction into three primary behaviours: consumption (viewing and reading), participation (interacting with content), and production (creating and uploading content). Following this framework, researchers have examined how engagement manifests across platforms, particularly differentiating between active participation and passive consumption. On Facebook and YouTube, active engagement involves actions such as liking, commenting, and sharing, whereas passive engagement consists of clicking, watching, or hovering over content

(Kaur et al., 2019; Khan, 2017). On X, active engagement further includes reposting, quoting, and following (Chen, 2011). Studies also highlight the prevalence of passive users (often called 'lurkers'), who primarily consume content without actively engaging, comprising up to 90% of users in many online communities (Nonnecke and Preece, 1999; Preece et al., 2004). This contrast between active contributors and passive consumers underscores the need to understand what motivates users to actively engage with content, particularly disinformation.

Due to the urgency of the topic, recent research has therefore explored the drivers of engagement with disinformation and fake news on social media. Emotionally charged content has been identified as one of the strongest amplifiers of engagement, with sensationalized headlines and narratives strategically crafted to trigger emotional responses such as fear, anger, and anticipation, thus encouraging interaction and dissemination (Horner et al., 2023). Additionally, visual elements seem to play a significant role in enhancing credibility and audience response (Cao et al., 2020; Viola, 2025a). Features such as clickbait, emotionally charged language, and references to specific individuals, organizations, or events further heighten emotional resonance, thereby boosting engagement (Ali et al., 2023). Other factors influencing the likelihood of sharing disinformation would include fear of missing out, source credibility, information quality, cognitive overload, and social media fatigue (Kumar et al., 2020; Islam et al., 2020). These content strategies would capitalize on psychological triggers to grab attention, manipulate perceptions, and enhance virality. This study builds upon these findings and integrates the Social Identity Theory (SIT) (Tajfel and Turner, 2004) with the SIMCA model (Turner, 1991; Tajfel and Turner, 2004; van Zomeren et al., 2008) to provide a theoretical lens for understanding engagement with disinformation.

3 Methodology

This study integrates SIT (Tajfel and Turner, 2004) with the SIMCA model (van Zomeren et al., 2008) to provide a theoretical lens for testing engagement with disinformation. Rather than accepting these models as given, we use them to formulate competing hypotheses regarding the drivers of user engagement. Engagement is operationalized as the cumulative sum of likes, shares, reposts, quotes, and replies associated with each post, as found in the literature[1] (Chen, 2011).

SIT posits that individuals derive a significant portion of their self-concept from their perceived membership in social groups. In the context of this study, the "in-group" is operationalised as the linguistic community (i.e., Anglosphere vs. Hispanosphere). The author acknowledges that language is not a one-to-one substitute for cultural identity (Edwards, 2009). At the same time, however, in the context of transnational disinformation particularly on social media, language barriers often form the primary perimeter of information ecosystems. Evidence shows that language is not merely symbolic but also structural: it organizes who interacts with whom and, by extension, what information circulates within which audiences. With regard to Twitter specifically, Hale (2014) demonstrates that its connectivity is strongly stratified by language, with interaction patterns clustering within language communities and comparatively fewer ties crossing language boundaries; multilingual users can bridge communities, but they do not erase the underlying segmentation. Similarly, Eleta and Golbeck' work on multilingual Twitter (2014) finds that most discourse remains anchored within language-specific public confirming that interaction on platforms frequently aligns with language-defined communities. Crucially, SIT suggests that different groups may develop distinct norms of interaction. Therefore, high engagement levels within a specific linguistic community may not necessarily reflect the quality of the content, but rather a community-specific baseline of expressive responding (Wintterlin et al., 2023). This framework helps explain why distinct engagement cultures might emerge independent of specific mobilizing cues, helping us to test the valence of SIT for disinformation research.

The SIMCA framework on the other hand posits that collective action is driven by three psychological predictors: social identity, perceived injustice, and participatory efficacy (van Zomeren et al., 2008). Applied to disinformation, SIMCA implies that narratives framing mainstream institutions as corrupt (Injustice) or urging users to "wake up" (Efficacy) should trigger a transition from passive consumption to active distribution. If SIMCA is universally applicable to social media disinforma-

[1]While high engagement metrics can indicate both support and heated opposition, from the SIMCA theoretical perspective, both reactions indicate that the content successfully triggered an engagement response, regardless of valence.

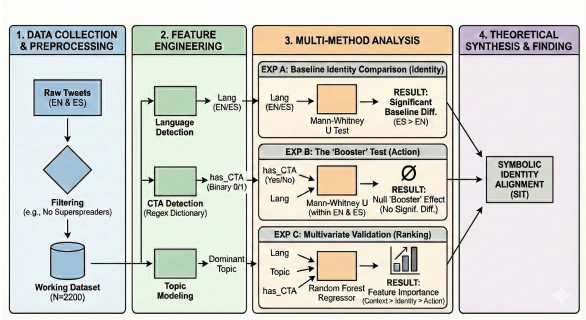

Figure 1: Research design: methodology overview.

tion, the presence of CA language should act as an engagement booster regardless of the language spoken, due to its explicit signalling of injustice and efficacy. By contrast, if engagement is symbolic, interaction may be driven primarily by the user's cultural baseline, with mobilizing language having little to no additive effect. We test this hypothesis using non-parametric statistical testing (Mann-Whitney U). To further validate the relative impact of mobilizing language compared to identity and contextual factors, we trained a Random Forest Regressor to predict engagement levels.

To identify CTA expressions without relying on a priori assumptions, we adopted a two-step 'Human-in-the-Loop' approach. We first applied Embedding-Based Topic Modelling using BERTopic (Grootendorst, 2022) to identify latent narrative clusters, the most driving topics, and a post excerpt representative of each topic. For each topic, we then qualitatively inspected the top 10 representative terms associated with mobilization. From these terms, we extracted a lexicon of explicit CTA expressions such as 'Resist the reset', 'donot-comply'), power-laden language (e.g., us vs. them frames and agency attribution such as 'Say no to digital IDs', 'Rise up!') and thematic organisation (e.g., New World Order, WEF Puppets). A post was coded as having CTA expressions only if it contained language from this manually validated lexicon. This ensured that the classification was interpretable and not subject to the noise of probabilistic topic modelling. Additionally, the author applied Critical Discourse Analysis (CDA) (Dijk, 1985, 1997) on a randomised sample of 200 posts (100 posts per language) to extract further expressions. The list of CTA expressions is provided in Table 4. A visual representation of the workflow is provided in Figure 1.

4 Data-set

Data retrieval was conducted through targeted queries that extracted posts containing specific hashtags, including *#Agenda2030*, *#The GreatReset*, *#NewWorldOrder*, *#wefpuppets*. These hashtags have been found in the literature as typically associated with disinformation, fake news, and conspiracy theories (Laquièze, 2025; Christensen and Au, 2023; Sa'ad Abdullahi and Pindiga, 2023). The full list of the seed hashtags is provided in the Appendix. Additionally, the data-set was enriched with information about content diffusion and filtered for posts drawn from sources known for low credibility, as identified in the Iffy+ list (Golding, 2025), which catalogues 2,042 outlets flagged by professional fact-checkers as non-reliable sources, a method commonly employed in prior work (DeVerna et al., 2024; Yang et al., 2021)[2].

The data-set was later pseudonymised to remove any identifiable references and it can be provided upon request to the author. The working data-set covers 394 days from 1 August 2022 to 30 August 2023. It contains 4,035 posts in two languages almost identically distributed (2,000 in English - EN, 2,035 in Spanish - ES) thus making the two sets highly comparable. The data-set also includes several attributes such as the post texts, the hashtags, likes, replies, reposts, shares, and quotes count. A detailed description is given in Table 7 in the Appendix.

5 Analysis

5.1 Superspreaders and bot activity

First we remove superspreaders and bot activity from the dataset. Superspreaders were identified using a multi-dimensional engagement and network-activity criterion, operationalized according to three criteria: per-user engagement metrics, conversational out-degree, and thresholding (DeVerna et al., 2024). Per-user engagement was calculated by computing three summary statistics on repost volume and three on reply volume for each user in the data-set as explained below:

- **Repost sum** ($\mathrm{rt\_sum}_i$): the total number of times user i's posts were reposted.

[2]The author acknowledges that while this strategy allows for broad coverage, it cannot account for the fact that not all the posts in the data-set contain links to external content and that individual articles from a low-credibility outlet may still be accurate.

- **Repost mean** ($\mathrm{rt\_mean}_i$): the average number of reposts per post for user i.
- **Repost max** ($\mathrm{rt\_max}_i$): the highest repost count received by any single post of user i.
- **Reply sum** ($\mathrm{rp\_sum}_i$): the total number of replies received by user i's posts.
- **Reply mean** ($\mathrm{rp\_mean}_i$): the average number of replies per post for user i.
- **Reply max** ($\mathrm{rp\_max}_i$): the maximum number of replies received by any single post of user i.

These six metrics captured both total volume (sum), typical activity level (mean), and individual peaks (max). Conversational out-degree was computed to assess how broadly each user initiated or contributed to discussions. This was operationalized by constructing a directed reply network in which each edge represented a reply from one user to another. A user's out-degree thus indicated the number of distinct users they replied to. Finally, the top five percent of users on each metric were identified independently to define the superspreader threshold. A user was designated a superspreader if they ranked in the top 5 percentile on any of the five measures (aggregate reposts, peak reposts, aggregate replies, peak replies, or out-degree). This approach ensured that both consistently active amplifiers and users responsible for individual cascades could be captured. The results stratified per language are displayed in Table 1 below.

Most notable insights include that among the identified superspreaders, English users (N = 248) tend to reply to more distinct users (median out-degree = 3) than Spanish users (median = 1). Spanish superspreaders receive slightly more replies on average (mean rp-sum = 7.57 vs. 6.73) and exhibit higher reply peaks (mean rp-x = 5.50 vs. 4.17). Repost behaviour is comparable across languages: mean repost peak (20) and sum (44 EN vs. 36 ES) show only modest differences. Based on the metrics, a total number of 1,915 superspreaders was identified (940 for English and 975 for Spanish). This means that in the English sub-set, superspreaders account for about 43% of posts, whereas in the Spanish sub-set, superspreaders contribute around 48%.

The remaining accounts were further inspected for potential bot activity using *BotometerLite* (Observatory on Social Media). *BotometerLite* produces scores derived from historical Twitter data collected prior to 31 May 2023, thereby aligning with the data-set's temporal scope and estimating the likelihood that a user was a bot at the time of observation. As none of the accounts was scored as a potential bot, no further filtering was applied to the data-set which finally included 1,140 posts in English and 1,060 in Spanish, making the two sets once again highly comparable.

5.2 Statistical analysis of language and engagement

The analysis now examines possible engagement differences between the two communities that can indicate that audience responses vary significantly according to community identity factors. To address the non-normal distribution of engagement metrics in social media data (Shapiro-Wilk $p < .001$), we employed the non-parametric Mann-Whitney U test. The results indicated a statistically significant difference in engagement between the language groups ($U = 558123.0, p < .001$). Although the median engagement for both groups was 0 (reflecting the long-tail nature of the data after removing superspreaders), the Spanish subset exhibited a higher mean engagement ($M = 1.87$) compared to the English subset ($M = 1.37$). The effect size was small ($r = 0.076$), suggesting that while language plays a statistically significant role in engagement patterns, it is likely one of multiple contributing factors. To ensure robustness, we repeated this analysis on the full dataset including superspreaders (Total N=4,035). The difference remained statistically significant ($U = 1,869,946, p < .001, r = 0.081$), confirming that the observed cross-linguistic difference is not an artifact of outlier exclusion. The results of both tests are displayed in Table 2. In the next section, the analysis tests the effect of CTA expressions in driving engagement with disinformation for both groups.

5.3 Collective action

This part of the analysis now investigates the role of CA language in user engagement with disinformation posts. The tested hypothesis is that perceived injustice strengthens motivation for collective action, which in an online environment would translate into engaging with and sharing disinformation as an act of defiance and regained agency. First, we identified CTA expressions combining the statistical (topic modelling) and linguistic (CDA) methodology as explained in 3.

Language	Statistic	out-degree	rp-max	rp-sum	rt-max	rt-sum
EN	50%	3.00	1.00	2.00	2.00	3.00
	75%	5.00	2.00	3.00	10.00	15.25
	Total	248.00	248.00	248.00	248.00	248.00
	Max	45.00	131.00	219.00	808.00	2495.00
	Mean	4.25	4.17	6.73	20.61	44.07
	Min	0.00	0.00	0.00	0.00	0.00
	Std	7.00	13.16	22.98	80.74	216.37
ES	50%	1.00	1.00	2.00	2.50	4.00
	75%	4.00	2.00	3.00	10.00	13.75
	Total	286.00	286.00	286.00	286.00	286.00
	Max	48.00	354.00	440.00	808.00	2495.00
	Mean	3.20	5.50	7.57	19.75	35.78
	Min	0.00	0.00	0.00	0.00	0.00
	Std	5.42	25.39	33.42	75.23	191.84

Table 1: Summary statistics (50th %, 75th %, total count, maximum, mean, minimum, and standard deviation) of superspreader metrics by language (EN vs. ES).

Group	N	Mean	Mdn	U-Stat	p	r
Dataset: Filtered (No Superspreaders)						
EN	1,140	1.37	0.0	558,123	$< .001$	.08
ES	1,060	1.87	0.0			
Dataset: Full (With Superspreaders)						
EN	2,000	8.72	1.0	1,869,946	$< .001$	.08
ES	2,035	14.19	1.0			

Table 2: Comparison of engagement metrics between English (EN) and Spanish (ES) groups with and without superspreaders

5.3.1 Topic modelling

Topic modelling was used to identify posts containing mobilising language and call-to-action expressions. Specifically, we used BERTopic with a multilingual SentenceTransformer (Grootendorst, 2022) as in the literature, it is found to outperform traditional LDA models on multilingual, short texts. Feature extraction was performed by applying a bag-of-words with unigrams and bigrams, and `max_df` $= 0.85$ and `min_df` $= 0.02$ to remove overly common/rare terms. Preliminary runs with unconstrained topic discovery produced a large number of fine-grained clusters, many of which overlapped semantically. Setting `nr_topics=10` allowed for a more coherent, high-level representation of discourse themes while retaining sufficient diversity to capture major narrative patterns across languages. This choice also ensured comparability across analyses and facilitated qualitative interpretation. The model returned per-document topic probabilities, per-topic term rankings, and most representative document per topic, enabling both quantitative summaries and qualitative interpretation across languages crucial for the extraction of CTA expressions.

Several topics identified by the model displayed strong mobilising language, encouraging resistance, participation, or direct action against perceived threats linked to Agenda2030 and related governance narratives. Representative posts (Table 3) show imperatives such as 'wake up', 'join' and 'resist', reflecting how conspiracy-linked discourses often blend moral urgency with collective calls to action. These patterns suggest that in disinformation discourse, mobilising language is often present as shared narratives of opposition and control.

Additionally, further expressions were extracted applying CDA (Dijk, 1985, 1997) on a sample of 200 posts (100 posts per language). The full list of expressions (91 in English and 82 in Spanish) is provided in Table 4.

6 Call-to-Action and user engagement

To test the hypothesis that mobilizing language functions as an engagement booster (as predicted by SIMCA), we now compared the engagement levels of posts containing CTA expressions against those without, within each language community. Results are displayed in Table 5. Contrary to theoretical expectations, no significant engagement premium was found for mobilizing content in either group. In the English data-set, posts with CTA expressions did not elicit significantly higher engagement than non-CTA posts ($U = 57,446, p = .39$). Similarly, in the Spanish data-set the presence of mobilizing language had no significant impact on engagement levels ($U = 23,054, p = .56$). These findings suggest that in social media disinformation ecosystems, the mere presence of explicit 'calls

Topic ID	Count	Top Representative Terms	Representative Example (excerpt)
0	1689	agenda2030, nwo, video, com, world, order, people, global, reset, truth	"Immortalized cell lines used in lab-grown meat... global elites want to control the food supply."
1	362	chile, españa, onu, méxico, argentina, política, gobierno, países, sociedad, latinoamérica	"Y ahora @GiorgioJackson, @gabrielboric... hablan de la ONU y la Agenda2030 en Latinoamérica."
2	292	climatescam, climate, agenda2030, globalwarming, wef, co2, fake, scam, science, propaganda	"Desmintiendo el calentamiento global bufones de la ONU... #ClimateScam."
3	46	canada, canadians, trudeau, agenda2030, climate, policy, protest, covid, freedom, rights	"Canadians need to support Alberta Oil & Gas — Trudeau is destroying our economy for Agenda2030."
4	37	vaccine, covid, depopulation, populationcontrol, health, agenda, wef, control, elite, world	"Vaccines are part of the depopulation agenda — open your eyes."
5	26	climatechange, greenagenda, sdgs, un, sustainability, policy, leaders, summit, goals, development	"The UN Climate Summit pushes Agenda2030 — same green agenda with new branding."
6	19	esg, economy, finance, corporations, capitalism, elites, governance, power, wef, global	"ESG is just a corporate version of Agenda2030 — economic control by the few."
7	16	energy, oil, gas, renewables, canada, europe, policy, crisis, cost, transition	"Energy transition in Europe is a scam — higher costs and less freedom."
8	14	digitalid, cbdc, surveillance, government, privacy, citizens, control, technology, freedom, rights	"Digital ID and CBDCs mean full surveillance — say no to control systems."
9	12	conspiracy, hoax, lies, fake, propaganda, media, agenda2030, truth, narrative, misinformation	"The media lies again — Agenda2030 is propaganda to hide the real plan."

Table 3: Topics identified by BERTopic with document count, most representative keywords, and a representative post excerpt.

EN	ES
wake, resist, join, sign, take action, act, must, do not comply, resistthereset, stop, share, help, demand, make, stand, stand up, speak out, raise your voice, get involved, take a stand, participate, march, protest, mobilize, defend, spread the word, volunteer, organize, boycott, take part, engage, rally, fight back, join the movement, act now, donotcomply	despierta, resiste, únete, firma la petición, actúa ahora, comparte, movilízate, haz algo, noalaagenda2030, noalplan, detente, toma, actúa, firma, ayuda, exige, haz, mantente, levántate, alza la voz, participa, actúa ya, protesta, lucha, defiende, propaga el mensaje, voluntario, organiza, boicotea, toma parte, únete al movimiento, comprométete, alístate, actúa hoy, únete ahora, haz tu voz escuchada, súmate, movilízate hoy

Table 4: Call-to-Action Expressions by Language

Lang	CTA	N	Mdn	U	p	r
EN	Yes	110	0.0	57,446	.39	-.01
	No	1030	0.0			
ES	Yes	46	0.0	23,054	.56	.01
	No	1014	1.0			

Table 5: Comparisons of engagement for posts with and without CTA expressions.

Feature	Category	Importance
Dominant Topic	Context	**43.8%**
Language (EN/ES)	Identity	**37.4%**
Collective Action	Mobilization	18.8%

Table 6: Feature Importance scores from the Random Forest Regressor predicting user engagement.

to fight' or 'wake up' does not trigger increased user interaction, reinforcing the interpretation that engagement is driven by community norms, e.g., identity rather than mobilizing appeals, i.e., action.

7 Multivariate Analysis: Drivers of Engagement

To further investigate the drivers of engagement, we trained a Random Forest Regressor using Language (Identity), Dominant Topic (Context), and Collective Action (Mobilization) as features.The model yielded a negative R^2 score (-0.003), indicating that textual and linguistic features alone cannot predict the magnitude of user engagement. This null result is however theoretically significant: it demonstrates that engagement in disinformation ecosystems is not a mechanical response to specific trigger words or topics, but likely a stochastic process driven by algorithmic amplification and network dynamics (Gonzalez-Bailon et al., 2011). At the same time, the Feature Importance analysis (Table 6 and Figure 2) provides insight into the relative weight of these weak signals. Dominant Topic (43.8%) and Language (37.4%) largely outperformed CTA expressions (18.8%). This ranking confirms that even among the weak textual predictors available, the context (Topic) and community (Language) carry twice as much weight as the mobilizing rhetoric (CTA).

8 Discussion

This study provided empirical evidence that challenges the universalist application of the SIMCA model to disinformation on social media. By disentangling the effects of community identity from

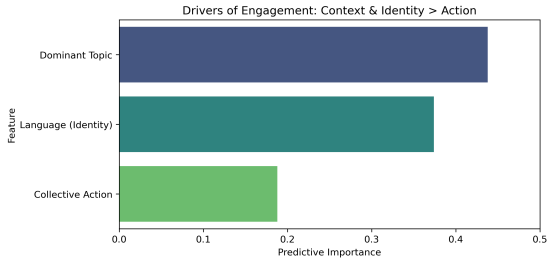

Figure 2: Feature importance ranking from the Random Forest model predicting user engagement.

mobilizing language (CTA expressions), our analysis revealed that language and cultural identity remain a primary driver of engagement. Cross-linguistic comparison between the two language communities showed a significantly higher baseline of activity in the Spanish-speaking community compared to the Anglosphere, a difference that persisted even after excluding superspreaders. This finding supports the Social Identity Theory perspective that online communities exhibit distinct engagement behaviours heavily structured by language (Hale, 2014).

Second, and most critically, our results demonstrate that mobilizing rhetoric fails to translate into measurable action such as explicit endorsement. Contrary to the predictions of the SIMCA model, the presence of explicit CTA language provided no significant engagement effect in either language group. This null result was further corroborated by multivariate analysis, where the CA feature proved to be the weakest predictor of engagement (18.8%), far outweighed by narrative context (43.8%) and language identity (37.4%).

These findings fundamentally reframe how we should interpret engagement metrics in disinformation research. The fact that users do not engage more with posts demanding them to act, resist, or wake up suggests that their interaction is symbolic rather than functional. In other words, users likely engage explicitly with a disinformation post to signal their alignment with the group and opposition to the mainstream (identity performance), rather than exclusively signalling collective resistance.

These results encourage us to rethink current applications of collective action models to social media as they may overestimate the mobilizing nature of engagement. Digital interactions in this context should be viewed as low-cost expressive behaviours, that is forms of identity signalling that satisfy affective needs for belonging without necessarily implying a commitment to collective mobilization. Finally, by documenting the significant baseline disparity between Spanish and English communities, this study challenges the field's reliance on Anglocentric data. It demonstrates that engagement is a culturally situated behaviour, proving that theoretical models built solely on English datasets cannot be generalized to other linguistic ecosystems without local validation.

9 Conclusion

This study set out to test the validity of the Social Identity Model of Collective Action (SIMCA) to social media disinformation. The findings suggest that engagement with disinformation functions primarily as symbolic identity alignment rather than functional collective action. Although mobilizing language is pervasive in disinformation discourse, its presence does not trigger the behavioural response predicted by collective action models. Instead, users appear to engage with content to signal group belonging and opposition to the mainstream, regardless of whether the content explicitly demands action.

By refining the application of SIT and SIMCA to social media, this study highlights the need to reconsider how engagement metrics are interpreted in disinformation research. While collective action models remain valuable for understanding how identities and grievances form, their explanatory power diminishes when mapped directly onto platform metrics. We argue that high engagement levels should be interpreted as signals of community cohesion and future research should therefore explore how cultural identity, emotional valence, narrative framing, and platform affordances shape identity-driven engagement with disinformation to develop more effective analytical frameworks and intervention strategies against the spread of disinformation online.

10 Limitations

While this study provides valuable insights, several limitations must be acknowledged. The data-set consists of posts from X, which may not be representative of broader social media engagement patterns on Facebook, YouTube, or TikTok. Future studies could investigate how different platforms' algorithmic amplification patterns alter engagement dynamics. Due to lack of resources, the study primarily analysed explicit collective action

expressions from a sample of the data-set but did not account for *all* the mobilization strategies in the data-set. Further research should incorporate larger semantic and qualitative analysis to assess a larger range of such linguistic strategies.

Declaration on Generative AI

During the preparation of this work, the author used Gemini 3 in order to: Grammar and spelling check. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

Séraphin Alava, Divina Frau-Meigs, Ghayda Hassan, Hasna Hussein, and Yuanyuan Wei. 2017. *Youth and violent extremism on social media: violent extremism on social media: mapping the research*. United Nations Educational, Scientific, and Cultural Organization, Paris. OCLC: 1089113298.

Maged Ali, Lucas Moreira Gomes, Nahed Azab, João Gabriel de Moraes Souza, M. Karim Sorour, and Herbert Kimura. 2023. Panic buying and fake news in urban vs. rural England: A case study of twitter during COVID-19. *Technological Forecasting and Social Change*, 193:122598.

Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236.

Carolin Amlinger. 2022. *Gekränkte Freiheit: Aspekte des libertären Autoritarismus*, erste auflage, originalausgabe edition. Suhrkamp, Berlin.

Shelly Ghai Bajaj. 2024. Digital Disinformation Threats and Ethnocultural Diasporas. In Gitanjali Adlakha-Hutcheon and Candyce Kelshall, editors, *(In)Security: Identifying the Invisible Disruptors of Security*, pages 53–65. Springer Nature Switzerland, Cham.

Zach Bastick. 2021. Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. *Computers in Human Behavior*, 116:106633.

Zack Beauchamp. 2019. Social media is rotting democracy from within. *Vox*.

Clare Birchall and Peter Knight. 2022. *Conspiracy Theories in the Time of Covid-19*, 1 edition. Routledge, London.

Erika Bonnevie, Allison Gallegos-Jeffrey, Jaclyn Goldbarg, Brian Byrd, and Joseph Smyser. 2021. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. *Journal of communication in healthcare*, 14(1):12–19. ISBN: 1753-8068 Publisher: Taylor & Francis.

Roderick J. Brodie, Ana Ilic, Biljana Juric, and Linda Hollebeek. 2013. Consumer engagement in a virtual brand community: An exploratory analysis. *Journal of Business Research*, 66(1):105–114.

Suzanne Brunsting and Tom Postmes. 2002. Social Movement Participation in the Digital Age: Predicting Offline and Online Collective Action. *Small Group Research*, 33(5):525–554. Publisher: SAGE Publications Inc.

Michael Butter and Peter Knight, editors. 2020. *Routledge handbook of conspiracy theories*. Routledge, Abingdon, Oxon ; New York, NY.

Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the Role of Visual Content in Fake News Detection. In Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu, editors, *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 141–161. Springer International Publishing, Cham.

Gina Masullo Chen. 2011. Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others. *Computers in Human Behavior*, 27(2):755–762.

Li Chen, Qi Ling, Tingjia Cao, and Ke Han. 2020. Mislabeled, fragmented, and conspiracy-driven: a content analysis of the social media discourse about the HPV vaccine in China. *Asian Journal of Communication*, 30(6):450–469.

Li Chen, Yafei Zhang, Rachel Young, Xianwei Wu, and Ge Zhu. 2021. Effects of Vaccine-Related Conspiracy Theories on Chinese Young Adults' Perceptions of the HPV Vaccine: An Experimental Study. *Health Communication*, 36(11):1343–1353.

Michael Christensen and Ashli Au. 2023. The great reset and the cultural boundaries of conspiracy theory. *International Journal of Communication*, 17:19–19.

Maura Conway. 2017. Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research. *Studies in Conflict & Terrorism*, 40(1):77–98. Publisher: Routledge _eprint: https://doi.org/10.1080/1057610X.2016.1157408.

Massimiliano Demata, Virginia Zorzi, and Angela Zottola, editors. 2022. *Conspiracy theory discourses*. Number volume 98 in Discourse approaches to politics, society and culture. John Benjamins Publishing Company, Amsterdam ; Philadelphia.

Matthew R. DeVerna, Rachith Aiyappa, Diogo Pacheco, John Bryden, and Filippo Menczer. 2024. Identifying and characterizing superspreaders of low-credibility content on Twitter. *PLOS ONE*, 19(5):e0302201.

Teun A. van Dijk, editor. 1985. *Discourse and Communication: New Approaches to the Analysis of Mass Media Discourse and Communication*. DE GRUYTER.

Teun Adrianus van Dijk. 1997. *Discourse studies: a multidisciplinary introduction. 2, 2,*. Sage. OCLC: 634170329.

John Edwards. 2009. *Language and Identity: An introduction*, 1 edition. Cambridge University Press.

Irene Eleta and Jennifer Golbeck. 2014. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, 41:424–432.

Don Fallis. 2009. A Conceptual Analysis of Disinformation.

Barrett Golding. 2025. Iffy Index of Unreliable Sources.

Sandra Gonzalez-Bailon, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. 2011. The Dynamics of Protest Recruitment through an Online Network. *Scientific Reports*, 1(1):197. ArXiv:1111.5595 [physics].

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Keith Gunaratne, Eric A. Coomes, and Hourmazd Haghbayan. 2019. Temporal trends in anti-vaccine discourse on Twitter. *Vaccine*, 37(35):4867–4871.

Scott A. Hale. 2014. Global connectivity and multilinguals in the Twitter network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 833–842, Toronto Ontario Canada. ACM.

Christy Galletta Horner, Dennis Galletta, Jennifer Crawford, and Abhijeet Shirsat. 2023. Emotions: The Unexplored Fuel of Fake News on Social Media. In *Fake News on the Internet*. Routledge. Num Pages: 28.

A. K. M. Najmul Islam, Samuli Laato, Shamim Talukder, and Erkki Sutinen. 2020. Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective. *Technological Forecasting and Social Change*, 159:120201.

Jessica Johnson. 2018. The Self-Radicalization of White Men: "Fake News" and the Affective Networking of Paranoia. *Communication, Culture and Critique*, 11(1):100–115.

Wandeep Kaur, Vimala Balakrishnan, Omer Rana, and Ajantha Sinniah. 2019. Liking, sharing, commenting and reacting on Facebook: User behaviors' impact on sentiment intensity. *Telematics and Informatics*, 39:25–36.

M. Laeeq Khan. 2017. Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior*, 66:236–247.

Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27.

Gaurav Kumar, Rishabh Joshi, Jaspreet Singh, and Promod Yenigalla. 2020. AMUSED: A Multi-Stream Vector Representation Method for Use in Natural Dialogue. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 750–758, Marseille, France. European Language Resources Association.

Ugo Laquièze. 2025. #LGBTpropaganda #GenderTheory #Wokism: Expanding and blurring the boundaries of francophone anti-gender discourse propagated on Twitter. *Politikon: The IAPSS Journal of Political Science*, 59(1):88–114.

Becca Lewis and Alice E. Marwick. 2017. Media Manipulation and Disinformation Online. Technical report, Data & Society. Publisher: Data & Society Research Institute.

Neil MacFarquhar. 2016. A Powerful Russian Weapon: The Spread of False Stories - The New York Times.

Guri Nordtorp Mølmen and Jacob Aasland Ravndal. 2021. Mechanisms of online radicalisation: how the internet affects the radicalisation of extreme-right lone actor terrorists. *Behavioral Sciences of Terrorism and Political Aggression*, 0(0):1–25. Publisher: Routledge _eprint: https://doi.org/10.1080/19434472.2021.1993302.

Blair Nonnecke and Jennifer Preece. 1999. Shedding light on lurkers in online communities. *Ethnographic studies in real and virtual environments: Inhabited information spaces and connected communities, Edinburgh*, 123128.

Observatory on Social Media. Botometer X.

Jenny Preece, Blair Nonnecke, and Dorine Andrews. 2004. The top five reasons for lurking: improving community experiences for everyone. *Computers in human behavior*, 20(2):201–223. Publisher: Elsevier.

Ignacio Ojea Quintana, Ritsaart Reimann, Marc Cheong, Mark Alfano, and Colin Klein. 2022. Polarization and trust in the evolution of vaccine discourse on Twitter during COVID-19. *PLOS ONE*, 17(12):e0277292. Publisher: Public Library of Science.

Madhavi Reddi, Rachel Kuo, and Daniel Kreiss. 2023. Identity propaganda: Racial narratives and disinformation. *New Media & Society*, 25(8):2201–2218. Publisher: SAGE Publications.

Sven Reichardt, editor. 2022. *Die Misstrauensgemeinschaft der "Querdenker": die Corona-Proteste aus*

kultur- und sozialwissenschaftlicher Perspektive, sonderausgabe für die bundeszentrale für politische bildung edition. Number Band 10857 in Schriftenreihe. Bundeszentrale für Politische Bildung, Bonn.

Stephen D. Reicher and Alexander S. Haslam. 2016. Fueling Terror: How Extremists Are Made. *Scientific American*.

Bashir Sa'ad Abdullahi and Habeeb Idris Pindiga. 2023. Tracking the Diffusion of Disinformation on the SDGs Across Social Media Platforms. In Jan Servaes and Muhammad Jameel Yusha'u, editors, *SDG18 Communication for All, Volume 2: Regional Perspectives and Special Cases*, pages 145–174. Springer International Publishing, Cham.

Hyunjin Seo and Robert Faris. 2021. Comparative Approaches to Mis/Disinformation| Introduction. *International Journal of Communication*, 15(0):8. Number: 0.

Guosong Shao. 2009. Understanding the appeal of user-generated media: a uses and gratification perspective. *Internet Research*, 19(1):7–25. Publisher: Emerald Group Publishing Limited.

Kate T. Simms, Sharon J. B. Hanley, Megan A. Smith, Adam Keane, and Karen Canfell. 2020. Impact of HPV vaccine hesitancy on cervical cancer in Japan: a modelling study. *The Lancet. Public Health*, 5(4):e223–e234.

Bonnie Stabile, Aubrey Grant, Hemant Purohit, and Kelsey Harris. 2019. Sex, Lies, and Stereotypes: Gendered Implications of Fake News for Women in Politics. *Public Integrity*, 21(5):491–502. Publisher: Routledge _eprint: https://doi.org/10.1080/10999922.2019.1626695.

Henri Tajfel and John C. Turner. 2004. *The Social Identity Theory of Intergroup Behavior*. Political psychology: Key readings. Psychology Press, New York, NY, US. Pages: 293.

Joshua A. Tucker, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature.

John C. Turner. 1991. *Social influence*. Social influence. Thomson Brooks/Cole Publishing Co, Belmont, CA, US. Pages: xvi, 206.

Jan-Willem Van Prooijen and Mark Van Vugt. 2018. Conspiracy Theories: Evolved Functions and Psychological Mechanisms. *Perspectives on Psychological Science*, 13(6):770–788.

Jan-Willem Van Prooijen and Karen M. Douglas. 2018. Belief in conspiracy theories: Basic principles of an emerging research domain. *European Journal of Social Psychology*, 48(7):897–908.

Martijn van Zomeren, Tom Postmes, and Russell Spears. 2008. Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives. *Psychological Bulletin*, 134(4):504–535. Place: US Publisher: American Psychological Association.

Lorella Viola. 2025a. What about Language? A Multilingual Behavioural Study of User Engagement with Disinformation on X: 5th Workshop on Reducing Online Misinformation through Credible Information Retrieval, ROMCIR 2025. *ROMCIR 2025 Reducing Online Misinformation through Credible Information Retrieval 2025*, pages 54–69. Publisher: CEUR-WS.

Lorella Viola. 2025b. 'Barren lesbians plotting sterilization': gender stereotypes and prejudices in health disinformation narratives, a cross-cultural analysis of social media of the HPV vaccine. In Catherine Tebaldi, Alistair Plum, and Christoph Purschke, editors, *Conspiracy as Genre: Narrative, Power and Circulation*. Bloomsbury Academic, London.

Bradley Wiggins. 2023. 'Nothing Can Stop What's Coming': An analysis of the conspiracy theory discourse on 4chan's /Pol board. *Discourse & Society*, 34(3):381–398. Publisher: SAGE Publications Ltd.

Florian Wintterlin, Tim Schatto-Eckrodt, Lena Frischlich, Svenja Boberg, Felix Reer, and Thorsten Quandt. 2023. "It's us against them up there": Spreading online disinformation as populist collective action. *Computers in Human Behavior*, 146:107784.

Asami Yagi, Yutaka Ueda, and Tadashi Kimura. 2024. HPV Vaccine Issues in Japan: A review of our attempts to promote the HPV vaccine and to provide effective evaluation of the problem through social-medical and behavioral-economic perspectives. *Vaccine*, 42(22):125859.

Kai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. 2021. The COVID-19 Infodemic: Twitter versus Facebook. *Big Data & Society*, 8(1):20539517211013861.

A Appendix

Table 7: data-set description

Value	Count
Likes	26,007
Quotes	1,281
Replies	3,329
Reposts	15,693
Users	2,091
Mentions	1448
Hashtags	13,691

B Appendix: Seed Hashtags

Table 8: Frequency of seed hashtags in the dataset. The list is sorted by frequency.

Hashtag	Count	Hashtag	Count	Hashtag	Count
#Agenda2030	2,950	#chemtrails	44	#klausschwab	23
#agenda2030	336	#BillGates	42	#WEFpuppet	22
#WEF	291	#SocialCreditSystem	42	#noalaagenda2030	22
#ClimateScam	147	#PureBlood	40	#malditaagenda2030	22
#GreatReset	144	#WEFpuppets	38	#Fauci	22
#NWO	121	#billgates	34	#depopulation	21
#NewWorldOrder	112	#Globalist	32	#NetZero	21
#Agenda2030.	90	#ClimateCult	32	#DictaduraSanitaria	20
#AGENDA2030	82	#SDGs	30	#Soros	20
#NOM	76	#TheGreatReset	29	#CrimesAgainstHumanity	20
#nwo	68	#Bilderberg	28	#Agenda2030?	20
#Agenda2030,	66	#Repentinitis	28	#CBDCs	19
#Agenda21	66	#NuevoOrdenMundial	27	#vaccineinjuries	18
#Plandemia	65	#VaccineGenocide	27	#CIA	18
#DigitalID	61	#FBI	26	#agenda21	18
#ODS	58	#wef	26	#FueraONU	17
#KlausSchwab	57	#GeorgeSoros	26	#WorldEconomicForum	17
#DiedSuddenly	56	#vaccine	25	#Nuremberg2	16
#15minutecities	55	#plandemia	25	#Chemtrails	16
#CBDC	50	#ESG	25	#15MinuteCities	15
#DepopulationAgenda	50	#HunterBiden	24	#Jarnac	15
#climatescam	48	#WEF2030Agenda	24	#VaccineSideEffects	14
#cbdc	45	#repentinitis	23	#CambioClimatico	14