Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

The Diffusion of Causal Language in Social Networks

Zhuoyu Shi[1,2], **Fred Morstatter**[1,2]

[1]Thomas Lord Department of Computer Science, University of Southern California, USA
[2]Information Sciences Institute, University of Southern California, USA
zhuoyush@usc.edu, morstatt@usc.edu

Abstract

Causal reasoning plays a central role in human cognition. It facilitates the ability to infer, predict, and manipulate outcomes within the environment, which in turn lays the foundation for a uniquely adaptive decision-making framework that is crucial in navigating complex problem-solving contexts. With the pervasive influence of social media platforms, these online social networks have become critical for disseminating information, shaping public beliefs, and influencing daily life. However, no study has examined the propagation of causal language within social networks. In this work, we analyze the dispersion of messages containing causal language against those without, within the milieu of a large online social network. With the entirety of messages over one complete day on Twitter along with two additional days for validation, and with our validated ensemble method for identifying causal language, our findings reveal that messages with causal language exhibit a more extensive reach than those without. Furthermore, our counterfactual analysis demonstrates that the effect of causal language on information diffusion is truly *causal*. Moreover, our findings indicate that messages incorporating causal language manifest a higher ability to spread to out-groups compared to those without. These novel insights reveal the unique diffusion pattern of causal language within social networks, and suggest a potential to mitigate the echo chamber effect, while causal language could serve as a bridge for diverse perspectives.

Introduction

Causality, as a foundational tenet of human comprehension of natural phenomena, has maintained its relevance since antiquity. This concept was first articulated in the era of ancient Greece, where Aristotle notably proposed the paradigm of "Four Causes"–encompassing material, formal, efficient, and final dimensions–in an endeavor to delineate the underlying mechanisms governing events and phenomena, around 2300 years ago (Aristotle 350BCE). In more recent literature, causal inference has emerged as a crucial research focus, where causal claims are presented and analyzed from different perspectives (Holland 1986; Marini and Singer 1988; Pearl 2009; Pearl and Mackenzie 2018; Haber et al. 2018; Wright and Augenstein 2021).

Causal reasoning is crucial in human cognition, and serves as a critical framework for shaping our understanding of the natural world (Gopnik and Wellman 2012). Often regarded as a cognitive scaffold, the foundation of causality equips individuals with the essential tools needed to decipher and interpret the complexity and diversity of phenomena around us (Sloman 2005). Indeed, the integration of causality enables us to establish connections, find patterns, and ultimately craft a coherent knowledge structure (Murphy and Medin 1985). This structure, nurtured by causal relations, serves as a knowledge schema that guides our perception and thought processes, influencing both conscious decision-making and subconscious behavior (Murphy and Medin 1985). The indispensable role of causal information thus suggests that our cognitive development, learning methods, and even culture are profoundly influenced by the cause-and-effect principles omnipresent in our environment (Norenzayan and Heine 2005).

Despite a broad consensus within the scientific community regarding the integral role of causal information in human reasoning and in framing our understanding of the world (Gopnik and Wellman 2012; Sloman 2005; Murphy and Medin 1985; Norenzayan and Heine 2005), strikingly scant attention has been paid to the contagion of causal information. The human brain exhibits an innate propensity for perceiving cause-effect relationships, enhancing the appeal of causal language (Corrigan and Denton 1996). Additionally, causal explanations offer a scaffold for understanding complex phenomena, contributing to their greater transmissibility (Gopnik, Schulz, and Schulz 2007; Henrich 2001). The inherent structure of causal language facilitates narrative construction, a key component for engaging and memorable communication (Trabasso and Van Den Broek 1985). Furthermore, the predictive ability of causal information makes it a powerful tool for forecasting (Trabasso and Van Den Broek 1985).

In order to understand the role of causal language in the dissemination of information across social networks, we leverage the context of social media. Our focus is to understand how information diffusion is influenced by the use of causal language within social networks. This requires a complete set of social media posts within a time period. To address this need, we analyze a complete set of all messages within one day on Twitter (Pfeffer et al. 2023), and validate

Messages with Causal Language	Messages without Causal Language
Once you lose a customer **because of** bad service, it is almost impossible to get that customer back. **{LLM, KW}**	I am not "frail".**{}**
the mental image of joon in the library working on a project **causes** me such pain and stress like ohhhhh god. . . ,, working with him on a group project like i would rather d!e **{LLM, SRL}**	I love how my cousin Yesina baby girl has only met me like about 3 times and she loves me so much. She saw me this Saturday and she ran up to me and was screaming Tia vivi!!! **{}**
Bad habits **make** good memories **{LLM, SRL}**	TO THE MOOOOOON **{}**
i've maintained and i know that's rlly what i need **bc** if i lose much more i'm gonna end up in the ward but it's also **driving** me insane **{SRL, KW}**	he really likes eating carrots (he thinks they're cookies, dude!) **{}**
I'm growing so much as an artist right now **because of** my classes, it's **making** me so happy **{LLM, SRL, KW}**	DREAM IS DOING A FACE REVEAL????? ohhhhh im scared **{}**

Table 1: Examples of messages with or without causal language. Causal connections are in bold. Causal language refers to phrases, sentences, or discourse that demonstrate a cause-and-effect relationship between different components. For example, in the third example, "Bad habits make good memories" contains causal language, where there is a cause "bad habits" and an effect "good memories," connected by a causal connection "make." The {} following each message signifies the sub-methods that detect causal language.

our findings with two additional days. This comprehensive exploration aims to understand how causal language drives the diffusion of information within social networks.

In our study, we investigate the mechanisms underlying information diffusion within social networks, with a specific focus on the role of causal language. Previous studies have primarily focused on how causal reasoning works (Murphy and Medin 1985; Sloman 2005; Norenzayan and Heine 2005; Gopnik and Wellman 2012). In contrast, our research focuses on the distinct and crucial role of causal language in social transmission dynamics. To this end, we address several significant research questions related to the propagation of causal language, namely:

- **RQ1** - Does the presence of causal language increase information diffusion in social networks?
- **RQ2** - Is the effect of causal language on information diffusion truly *causal*?
- **RQ3** - How effectively do messages containing causal language spread to out-groups?

Initially, we develop an ensemble method with three distinct sub-methods for identifying the presence of causal language within messages, and then validate our method with well-trained human annotators. Then, we conduct analysis to understand the emotional and moral frameworks coexisting with causal language. Subsequently, we examine whether causal language manifests more extensive dissemination compared to non-causal language. We then conduct a counterfactual analysis to investigate whether the effect of employing causal language on information diffusion is truly *causal*. Lastly, we examine the contagion capacity of causal language, specifically in terms of its ability of spreading to out-groups. Our results suggest that causal language indeed leads to more dissemination than non-causal language in all aspects. These explorations are critical in both the context of contagion and broadly in communication phenomena, and suggest a potential to mitigate the echo chamber effect.

Dataset Description

The dataset used in this research is obtained from previous research that collected every message[1] within a 24 hour period from Twitter (Pfeffer et al. 2023). The collection covered a complete 24 hour window spanning from September 20, 15:00:00 UTC to September 21, 14:59:59 UTC in 2022, with 375 million posts in total. In this work, we only consider the diffusion of original (i.e., not a repost, reply, or quote post), English posts without hashtags or URLs, in order to achieve alignment with real-life social networks. We further validate our findings with data from two additional days, collected with the same methodology by the same authors (Pfeffer et al. 2023) on December 07, 11:00:00 UTC to December 08, 10:59:59 UTC in 2022, and January 25, 14:00:00 UTC to January 26, 13:59:59 UTC in 2023.

The initial analysis of message propagation within the 24-hour period reveals a notable observation: the majority of reposts predominantly occur within a brief, 21.24-minute window following the initial posting of a majority of original posts. Based on this insight, we remove all messages posted within the final hour of the day to ensure that all messages in our study have sufficient time to be reposted. As a result, our refined dataset comprises 6,598,454 original English posts from the preceding 23 hours of the referenced day. Notably, 561,245 of these were subject to at least one repost within the entire 24-hour period.

Detecting Causal Language

Similar to the method described in previous work on the spread of true and false news (Vosoughi, Roy, and Aral 2018), we employ an ensemble strategy that is composed of three distinct sub-methods for identifying the presence of causal language within messages. Our methodology utilizes a voting strategy, encompassing three discrete, yet inter-

[1] A message may also be referred to as a post, formerly referred to as a tweet. We use the terms "message" and "post" interchangeably in this section.

connected, techniques: a fine-tuned Large Language Model (LLM) (Liu et al. 2019; Priniski, Verma, and Morstatter 2023), a Semantic Role Labeling (SRL) method (Gardner et al. 2018; Wolff, Song, and Driscoll 2002), and a keyword-based method (Xu et al. 2020; Girju and Moldovan 2002). Under this system, a message is classified as containing causal language if and only if at least two out of the three sub-methods affirm its presence. Conversely, if no causal language is identified by any of the sub-methods, the message is deemed devoid of it. To ensure the robustness of our findings, messages that are marked by only one single sub-method are considered edge cases and excluded from subsequent analyses. The Venn diagram shown in Figure 1 illustrates the overlap among the three sub-methods employed in our ensemble approach, underscoring that each sub-method within our ensemble approach is indispensable. We hired annotators from Prolific to label a subset of our dataset, with each message being assessed by three separate annotators to ensure reliability. Via our ensemble detection method, messages containing causal language achieve a precision rate of 88%, whereas the messages without causal language reach a precision rate of 94%. This triadic detection system bolsters the accuracy of our process, facilitating a precise and reliable identification of messages that possess causal language. Our ensemble method is capable of capturing the semantics of the text, which goes beyond mere keyword-based approaches (Girju and Moldovan 2002). We have 5,935,862 messages remaining after this step, where 53,359 messages contain causal language, while 5,882,503 do not (see Table 1 for examples).

Fine-tuned Large Language Model Method

We utilize a state-of-the-art deep learning model (Priniski, Verma, and Morstatter 2023), which has been fine-tuned based on RoBERTa (Liu et al. 2019), a large language model. It not only discerns the presence or absence of a causal relationship within a given text, but also labels cause and effect span pairs. The automated identification and labeling abilities of the model enhance the precision of causal relationship detection, thus serving as a significant component of our analysis methodology.

Semantic Role Labeling Method

Semantic Role Labeling (SRL) is a computational method to identify an event (who did what to whom) mentioned in the text, and find the relationships between noun arguments connected by verbs at the sentence level. We utilize AllenNLP's tool (Gardner et al. 2018), a renowned Python library for natural language processing. It enables us to label semantic roles as proto-agent (serving as the cause) and proto-patient (serving as the effect), along with their connecting verbs in the text. Our approach integrates a list of causal verbs (Wolff, Song, and Driscoll 2002), from which we eliminate words "get," "set," and "have" due to their frequent non-causal usage in common language. A message is identified as containing causal language by the SRL method if it features proto-agent and proto-patient pair(s) whose connecting verb is a causal verb.

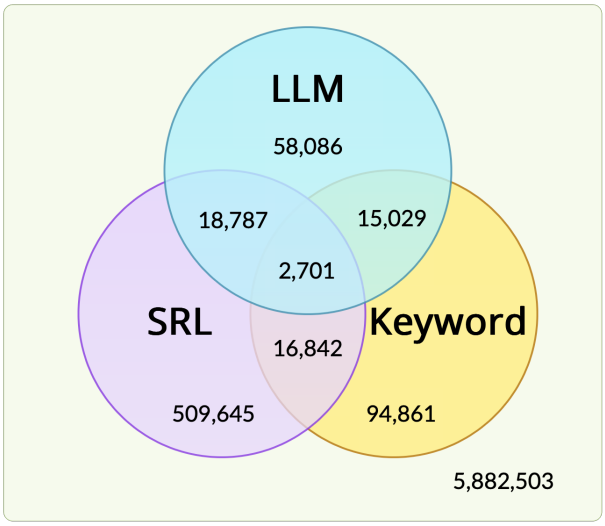


Figure 1: The Venn diagram illustrates the overlap of the three sub-methods used in our ensemble method.

Keyword-based Method

Lastly, we employ a classic approach that identifies the presence of causal language in text based on the existence of specific causal keywords. To achieve this, we use the keyword lists from previous studies (Xu et al. 2020; Girju and Moldovan 2002). In order to enhance the accuracy of our identification process, we exclude those keywords from our list that overlap with the causal verbs used in the Semantic Role Labeling (SRL) method. This keyword-based method, with its straightforward approach, contributing to the robustness of our ensemble strategy in detecting causal language.

Validation

To validate our detection results, we hired annotators to label a total of 300 messages, employing a random sampling strategy to select these messages. Specifically, 100 messages were randomly selected from each of the following three categories: the Causal Language (CL) group, where at least two out of three sub-methods confirmed the presence of causal language; the non-CL group, where none of the sub-methods detected causal language; and the edge case group, where exactly one sub-method indicated causal language. Annotators were hired from Prolific. Each annotator was trained with four examples with explanations, and then tested with six straightforward examples. We only included annotators who correctly annotated at least five out of six of these examples. Each annotator was tasked with annotating 30 messages and compensated $5 for their effort. Each message was labeled as "contains CL" or "does not contain CL" by three distinct annotators to ensure reliability. For an example of an annotation, please refer to Figure 6 in the Appendix.

The precision achieved by the CL group is 88%, while the non-CL group attains a precision of 94%. Among all annotated messages, there is a total of 100 messages that are marked as containing causal language using our ensemble

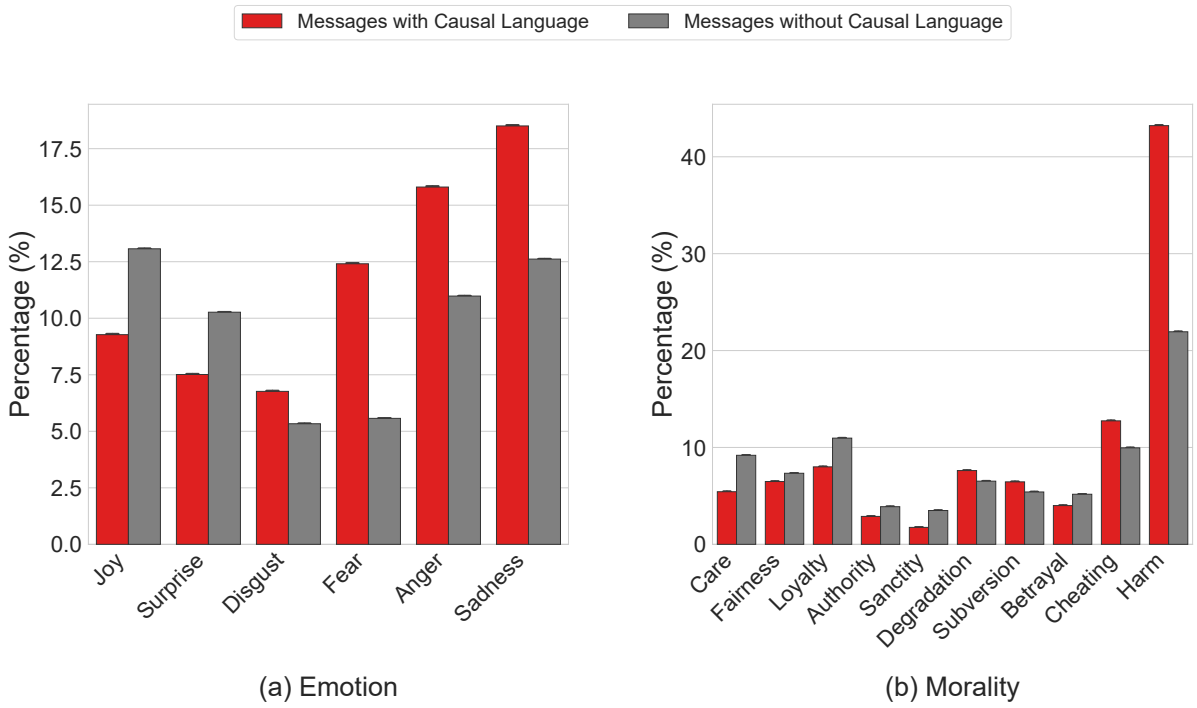


Figure 2: Comparative analysis of the percentage distribution of dominant (a) emotion or (b) morality categories in messages, segregated by the presence or absence of causal language. Error bars in both (a) and (b) reflect the variability across 100 bootstrapping runs. There is a consistently higher prevalence of negative emotion and moral expressions in messages containing causal language compared to those without.

method. The breakdown of these messages is as follows: 35 messages identified by the combination of {LLM, SRL}, 29 messages by {LLM, KW}, 31 messages by {SRL, KW}, and 5 messages by {LLM, SRL, KW}. The corresponding precision rates for these combinations are at 94.3%, 86.2%, 83.9%, and 80.0%, respectively. Similarly, there is a total of 100 messages that are marked as edge cases using our ensemble method. The breakdown of these messages across different methods is as follows: 8 messages identified by LLM, 77 messages by SRL, and 15 messages by KW. The corresponding precision rates for these distinct methods are 37.5%, 26.0%, and 6.7%, respectively. Please note that within each of the three groups annotated–the CL group, the non-CL group, and the edge case group–the distribution of subcategories in our annotated sample of 100 messages per group reflects their respective proportions in the entire dataset. This alignment with the overall message distribution has resulted in small sample sizes for some specific sub-groups, e.g., the {LLM, SRL, KW} combination in the CL group, LLM in the edge case group, etc. To further evaluate the consistency among annotators, we calculate the inter-annotator agreement using Fleiss' Kappa. The overall agreement is found to be 73.33% for the CL group, 84.00% for the non-CL group, and 70.67% for the edge case group. This indicates that every combination of sub-methods in our ensemble approach is indispensable. Our method is more accurate than labeling a message as containing CL based on only one sub-method.

Properties of Text That Co-Occur with Causal Language

First, we explore the textual characteristics that co-occur with the use of causal language. The critical role that emotion and morality play in defining our everyday existence has been acknowledged (Berger and Milkman 2012; Brady et al. 2017), revealing a strong connection with language. They provide a contextual framework for our experiences and interpersonal engagements, thereby enhancing our comprehension of others, swaying our ethical assessments, and steering our conduct across a multitude of contexts (Ekman 1999; Hofmann et al. 2014; Haidt 2001). These studies demonstrate the profound influence that emotion and morality have on the use of language, providing compelling insights into the complicated interplay of language, cognition, and social factors in communication.

In accordance with the Basic Emotion Theory postulated by Ekman (1999), we adopt a framework built around six foundational emotions: fear, anger, joy, sadness, disgust, and surprise. These emotions provide the bedrock upon which complex emotions can be understood. We adopt a state-of-the-art computational approach (Hartmann 2022) that leverages the strengths of large language models to identify emotions expressed in text. This approach is employed to dissect the emotional content of each message along seven distinct dimensions – namely fear, anger, joy, sadness, disgust, surprise, and a category reserved for non-emotional content.

When it comes to assessing the moral dimensions of our data, we turn to Moral Foundations Theory (Graham et al. 2013), which presents a robust framework that is grounded in five foundational pillars: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation. To measure moral content, we employ the Extended Moral Foundations Dictionary (eMFD), a lexicon-based method (Hopp et al. 2021). This methodology contextualizes each message within a ten-dimensional moral sphere encompassing the positive or negative of five moral foundations. Each dimension serves as a distinct yardstick, enabling us to map the complex moral landscape of the digital discourse.

Every message is categorized under at most one of six emotion labels. It is possible that it is identified as devoid of emotional content, in which case no emotion is assigned. Similarly, this applies to the assessment of morality, where each message either aligns with one of the ten moral dimensions, or is deemed to lack any moral content. The results shown in Figure 2(a) demonstrate a significant co-occurrence between the usage of causal language in messages and the emotional valence conveyed therein. The majority of messages employing causal language are found to express predominantly negative emotions such as sadness, anger, fear, and disgust. Conversely, the presence of positive emotions such as joy, is markedly lower in the messages with causal language. Shifting focus to the moral foundations underpinning these messages, Figure 2(b) provides a comparative analysis of those with causal language against those without. Messages containing causal language consistently exhibit a higher prevalence of expressions of negative morality. The interplay of causal language, emotion, and morality underscores the association of linguistic choices on the conveyance of emotional and moral sentiments.

These results mirror earlier investigations performed in controlled lab environments, where a marked propensity for individuals to utilize causal language when articulating negative events was noted (Rozin and Royzman 2001). This pattern appears to arise from a complex interplay of cognitive and psychological processes, including but not limited to coping mechanisms, an innate need for control, and a noticeable negativity bias (Lazarus and Folkman 1984; Thompson, Armstrong, and Thomas 1998; Rozin and Royzman 2001).

Measuring Attention to Causal Language

In order to understand whether the presence of causal language increases information diffusion in social networks, we use the number of reposts as a proxy for attention. We utilize percentiles as a method of representation in the analysis of repost count distributions of messages employing causal language versus non-causal language. Percentiles are robust to outliers, which are frequently encountered in social media datasets due to the presence of outliers that can drastically skew the mean and standard deviation, thus potentially offering a misleading representation of the data distribution. To control for outliers, we examine repost count distributions up to the 99th percentile as shown in Figure 3, rather than extending the analysis to the full 100%. By employing percentiles, we can better understand the relative standing of individual observations, giving a clearer insight into general repost behaviors without being heavily influenced by extreme values. Furthermore, utilizing percentiles allows for a straightforward comparison between different distributions, enabling us to effectively contrast the repost dynamics of two groups: messages employing causal language versus non-causal language.

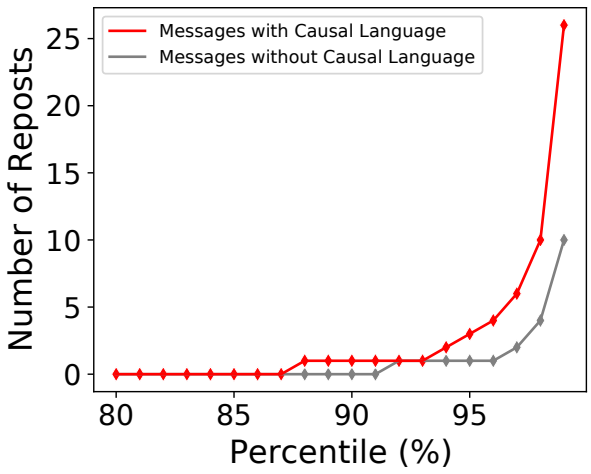


Figure 3: Comparison of repost count distribution for messages employing causal versus non-causal language. The percentile distribution for messages using causal language exhibits a significantly larger reach compared to those without causal language. Specifically, at the 99th percentile, messages with causal language achieve a repost count of 26. In contrast, for messages without causal language, the 99th percentile is 10.

Our findings, as illustrated in Figure 3, indicate a noteworthy trend in the diffusion of messages containing causal language against those without. Messages incorporating causal language are found to diffuse significantly more broadly than those devoid of such linguistic features. In order to validate these findings, we employ a Kolmogorov-Smirnov (K-S) test, a non-parametric method, to compare the distributions of repost count of messages with and without causal language. The K-S test confirms that the differences observed in the diffusion patterns of the two groups are statistically significant ($p < 10^{-5}$). This indicates that the presence of causal language increases information diffusion in social networks.

Counterfactual Analysis

We find that messages containing causal language receive a higher volume of reposts. However, it is possible that this is due to the presence of other confounds: moral language, the popularity of the author posting the message, among others. In this section, we investigate whether the effect of causal language on information diffusion is truly *causal*. We formulate this as a counterfactual design. We assemble pairs of messages that are as similar as possible, with the only difference being that one message in the pair contains causal

language and the other does not. We begin by creating a "treatment group" of messages containing causal language. Similarly, we create a "control group" of messages without causal language. For a fair comparison, messages in the control group should be as similar as possible to their counterparts in the treatment group, and both groups should be of equal size (Stuart 2010). Each message in the treatment group is paired with the most similar message without causal language, and these matched non-causal language messages then form the control group.

Selecting Covariates for Similarity Measurement

First, we select covariates for measuring similarity between a message with causal language and a message without causal language. We have the following covariates: emotion category, morality category, the author's follower count, the daily posting frequency of the author, and the semantic meaning. We select these covariates based on previous studies. Previous work has shown that number of followers and posting frequency are correlated with repost count (Webberley, Allen, and Whitaker 2011; Suh et al. 2010). Prior work also suggests text features, especially emotion and morality, are highly associated with information diffusion (Stieglitz and Dang-Xuan 2013; Heath, Bell, and Sternberg 2001; Brady et al. 2017).

Matching Method

We leverage a matching method to get the control group following Stuart (2010). For each message containing causal language, we get its emotion category, morality category, follower count category, daily posting frequency category, and embedded text vector of the message. These are computed as follows:

Emotion There are 7 categories in emotion: Joy, Anger, Disgust, Fear, Sadness, Surprise, and Non-emotion, as we compute and describe in "Properties of Text that Co-Occur with Causal Language."

Morality There are 11 categories in morality: Care, Harm, Fairness, Cheating, Loyalty, Betrayal, Authority, Subversion, Sanctity, Degradation, and Non-moral, as we compute and describe in "Properties of Text that Co-Occur with Causal Language."

Follower Count There are 5 categories, where n denotes follower count for each user: low ($0 \leq n \leq 10$), mid-low ($10 < n \leq 100$), mid ($100 < n \leq 1000$), mid-high ($1000 < n \leq 10000$), high ($10000 < n$).

Daily Posting Frequency There are 5 categories, where n denotes daily posting frequency of each user: low ($0 \leq n \leq 5$), mid-low ($5 < n \leq 10$), mid ($10 < n \leq 50$), mid-high ($50 < n \leq 100$), high ($100 < n$).

Text Embedding We compute an embedded text vector for each message with Sentence Transformers provided by Hugging Face (Reimers and Gurevych 2019).

With these covariates, we match each message from the treatment group to the non-causal language group to identify pairs. To ensure a valid match, we require all four covariate categories within one pair to be identical: emotion category, morality category, followers count category, and daily posting frequency category. Once these criteria are met, we compute the cosine similarity between the embedded text vector of each message in the treatment group and all eligible candidates in the non-causal language group. We then select the message with the highest cosine similarity from the non-causal language group to pair with each message in the treatment group, thereby forming the control group.

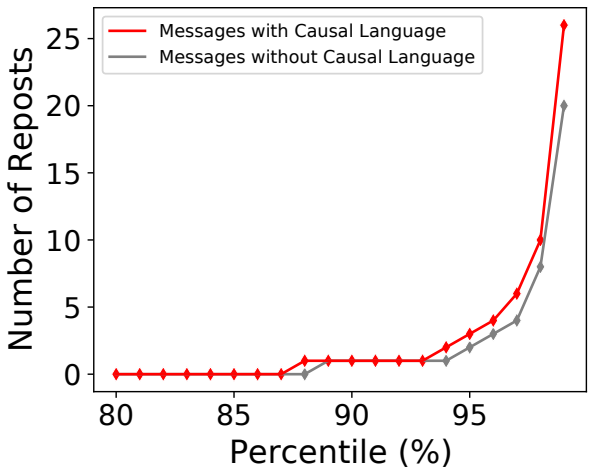


Figure 4: Comparison of repost count distribution for paired messages employing causal versus non-causal language via counterfactual analysis.

Counterfactual Results

Our matching approach yields 53,348 matched pairs (see Table 2 for examples). This means that 99.98% of the messages with causal language are matched to a message without causal language. Following the same methodology we used to answer RQ1, we measure the repost by percentile and employ the Kolmogorov-Smirnov (K-S) test. As shown in Figure 4, our findings reveal that the messages containing causal language have more reposts than those without causal language. The K-S test confirms that the differences observed in the diffusion patterns between the treatment and control groups, which are of equal size, are statistically significant ($p = 0.00070$). This suggests that the effect of causal language on information diffusion is truly causal.

Out-Group Diffusion of Causal Language

Then, we investigate the propagation of both causal and non-causal language and to out-groups, examining the effects in conjunction with emotion and morality. Employing the Louvain algorithm (Blondel et al. 2008) for community detection, we dissect the repost network drawn from the original dataset (Pfeffer et al. 2023). The use of repost network enables us to closely examine the dissemination of information and ideas within and across social groups on a large scale (Stieglitz and Dang-Xuan 2012). Consequently, each com-

	Treatment Group	Control Group
Example 1	Many songs of the imagine dragons **make** me cry	u know it's bad when you're crying to every song in folklore
Example 2	Styling myself for events has been so hard lately, **due to** this unpredictable weather	I don't know how to dress kinda weather
Example 3	I know its hard for people to understand each other. You think im crazy **because of** loneliness, but the way of life is to fo step by step. I know all about the lost love. Also worked hard to make up for you. Sometimes things go against our wishes.	I miss feeling calm and at peace. The days where I felt good with my life and so content with everything. Now I am just empty and broken, heart shattered and scarred, anxious and disturbed. And terribly missing you.

Table 2: Examples of pairs of messages with and without causal language via counterfactual analysis. Causal connections are in bold.

munity identified by the algorithm is construed as a "group." This comprehensive repost network is constructed by adding every author and reposter pair from the 24-hour dataset (Pfeffer et al. 2023) to the network, regardless of whether the original message fell within our designated 23-hour observation window. This approach results in a network comprising 11,729,159 nodes and 48,648,092 edges. For community detection, we employ the Louvain algorithm with 100 iterations, obtaining community detection results each time. Consider one iteration as an example: 196,830 communities were detected, with 5,585,019 inter-community edges, and 43,063,073 intra-community edges. This approach provides unique insight into the trajectory of causal and non-causal language diffusion, enhancing our understanding of the distinct patterns and behaviors associated with the spread of causal and non-causal language. The in-depth analysis of these propagation dynamics with or without causal language through the repost network helps in elucidating the nuances of social interactions and communication patterns that drive the spread of different types of language, thereby enriching our understanding of information flow within social networks.

We observe a distinct dichotomy in repost potential based on the presence or absence of causal language. Of the messages incorporating causal language, a noteworthy 12.7% are reposted, compared to 8.2% of those without causal language. This discrepancy also manifests when considering the diffusion of messages to out-group members. Specifically, 3.3% of messages that employ causal language effectively reach out-group audiences, compared to 1.7% when causal language is not used. These observations intimate that causal language has a greater potential for influence compared to non-causal language, with an enhanced ability to cross group boundaries.

To deepen our understanding, we disaggregate the results into finer categories, taking into consideration elements of emotion and morality. With Figure 5, we can compare differential patterns observed in the dissemination and out-group dissemination of messages based on the presence or absence of causal language. These patterns vary across different emotion categories. In analyzing the dissemination percentages, two distinct trends emerge. On the left of Figure 5, a comparative analysis between messages employing causal language and those without reveal noticeable variations across each distinct emotion category. The right of Figure 5 shows the dissemination to out-group members across the different emotion categories and demonstrates contrasts between messages with causal language versus those without. Remarkably, messages that include causal language consistently achieve higher levels of dissemination across all categories of emotion, with this effect being even more pronounced when considering the spread of messages to out-groups. Notably, these trends remain consistent even when dissected into morality categories, reinforcing the significant effect of causal language on message dissemination (see Appendix, Figure 7). These findings indicate the critical role of causal language in influencing the reach and impact of messages, which may have broader implications for understanding the mechanisms underlying social communication[2].

Validating Findings with Additional Data

To validate our findings, we analyze two additional complete days of Twitter data gathered using the same methodology on December 8, 2022, and January 26, 2023. In the dataset collected on December 8, 2022, there are 6,618,390 original English posts without hashtags or URLs from the first 23 hours of the day, and 517,083 of these were subject to at least one repost within the entire 24-hour period on that day. Among these, there are 55,086 messages containing causal language, while 5,896,374 do not. Similarly, in the dataset collected on January 26, 2023, there are 5,931,395 original English posts without hashtags or URLs from the first 23 hours of this day, and 474,707 of these have at least one repost within the entire 24-hour period of that day. There are 49,238 messages containing causal language, compared to 5,282,185 messages that do not. We find the results derived from these two datasets are nearly identical to those findings from our primary dataset (see Figures 8 - 11 in the Appendix). Importantly, the consistent outcomes across the different datasets underscore the robustness of our findings, providing further credibility and strength to our study.

[2]We also evaluate the findings in this section by implementing the Leiden algorithm (Traag, Waltman, and Van Eck 2019) for community detection, and observe the same trend presented in this section as we implement the Louvain algorithm.

Discussion and Conclusion

With one complete day of Twitter data along with two additional days for validation, and using our validated ensemble method for identifying causal language, we find that messages with causal language are reposted more than messages without. Our counterfactual analysis then reveals that the effect of causal language on information diffusion is truly *causal*. Furthermore, we also find that causal language is effective at spreading to out-groups. Additionally, we find that people tend to use more causal language when expressing negative emotions and morality, which is aligned with previous research conducted in lab environments (Heider 2013; Baumeister et al. 2001; Thompson 2016; Cacioppo and Gardner 1999; Rozin and Royzman 2001). These findings reveal how causal language diffuses in social networks.

We identify statistically significant variation in the diffusion patterns of messages containing causal language, thereby underscoring the critical need to scrutinize how distinct causal beliefs might be functionally interconnected to their respective modes of dissemination. Our analysis bifurcates the discourse into causal and non-causal language. Yet, it is plausible that causal language could be deconstructed into more nuanced subcategories. For instance, causal language can pivot on cultural or religious beliefs or revolve around empirical facts. These various subcategories might uniquely influence the course of social transmission. As such, future studies ought to elucidate the mechanisms through which different classifications of causal language motivate individuals to propagate and engage in discussions around their beliefs.

We discover that causal language is more effective at spreading to out-groups than non-causal language. The use of causal language could be a useful tool in addressing the echo chamber effect, where homogeneous opinions are amplified and reinforced while contradictory viewpoints are muffled in socio-informational spaces (Jamieson and Cappella 2008; Nyhan et al. 2023; González-Bailón et al. 2023). Echo chambers can reinforce misinformation, heighten polarization, and hinder the development of a comprehensive understanding of complex issues (Del Vicario et al. 2016). Causal language, through its capacity to clarify the cause-effect relationships of a particular topic, could facilitate the breaking down of these chambers (Sloman 2005). By articulating explanations in terms of causes and effects, it can help to shed light on otherwise complex and ambiguous phenomena, providing an avenue for constructive discussions based on logical reasoning rather than ideological beliefs (Mercier and Sperber 2017). Additionally, exposure to causal language could encourage individuals to engage with differing viewpoints and mitigating the echo chamber effect.

Our findings align with other previous research conducted in a lab environment, where causal language often co-occurs with negative emotions due to various cognitive and psychological processes. Attribution theory, for instance, posits that people seek explanations for events, especially negative ones, thus increasing the use of causal language (Heider 2013). Coping mechanisms can also trigger the use of causal language, as understanding the origins of negative emotions can help individuals manage their discomfort (Baumeister et al. 2001). A desire for control can similarly lead to increased use of causal language, as it may offer a sense of security and mastery over distressing events (Thompson 2016). The inherent complexity of negative emotions compared to positive ones can also result in more detailed processing, thereby provoking more causal thinking (Cacioppo and Gardner 1999). Lastly, negativity bias–a tendency to pay more attention to negative information–can also contribute to the increased use of causal language when processing negative experiences (Rozin and Royzman 2001).

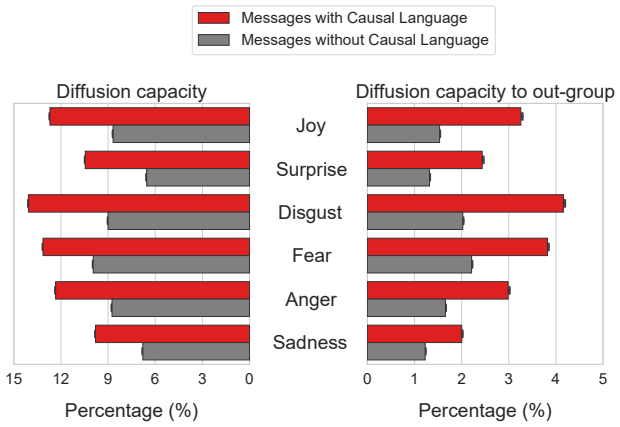


Figure 5: Differential dissemination patterns and out-group dissemination patterns of messages with or without causal language, across varied emotion categories. Error bars represent variability across 100 runs with the Louvain algorithm for community detection. (Left) Dissemination percentages for messages employing causal language versus non-causal language across each distinct emotion category. (Right) Dissemination percentages to out-group members across each different emotion category, for messages with causal language versus non-causal language. Messages with causal language generally result in higher dissemination and out-group dissemination, under all emotion subcategories.

We present a novel perspective on causal language research in psychology, drawing from empirical data. By analyzing real social networks we argue for a distinct advantage over laboratory-based investigations (Lazer et al. 2009). Engaging with these vast datasets facilitates researchers to access broader, more varied, and globally dispersed linguistic samples, amplifying the generalizability of insights and paving the way for firmer conclusions for studies of causal language (Mislove et al. 2011). Notably, this approach bypasses potential distortions of the Hawthorne Effect (Adair 1984), with naturally occurring language patterns offering genuine reflections free from observation-induced influences. Contrasting with laboratory environments where only fleeting linguistic behaviors might be captured, empirical data promotes longitudinal studies (Ruths and Pfeffer 2014), making it feasible to observe the diffusion of causal language use over durations. The inherent ecological validity of this method captures the intricacy of real-world linguistic behaviors and interactions (Back et al. 2010).

Related Work

The phenomenon of information diffusion in online social networks has been a topic of significant interest in recent research, engaging people in understanding how information spreads in social network platforms such as Twitter and Facebook. Various works have discussed the role of social networks in information diffusion and offered overviews of how information diffuses in online social networks (Berger and Milkman 2012; Guille et al. 2013; Li et al. 2017). Suh et al. conducted a large-scale analysis on factors that impact the repost rate on Twitter, shedding light on the essential parameters that control information diffusion on the platform (Suh et al. 2010). Berger and Milkman investigated what makes online content go viral, examining the attributes of the content that add to its virality (Berger and Milkman 2012). Brady et al. explored how emotion shapes the diffusion of moralized content in social networks (Brady et al. 2017). With the era of rampant online misinformation, Bakshy et al. analyzed rumor cascades (Friggeri et al. 2014), and Del Vicario et al. illustrated the spreading of misinformation in the digital space (Del Vicario et al. 2016). Vosoughi et al. distinguished between the spread of true and false news online (Vosoughi, Roy, and Aral 2018).

Limitations and Ethical Considerations

During the data collection period, the platform was known as Twitter, and messages were referred to as "tweets." Following the platform's rebranding to X, the term "tweet" has been updated to "post," in line with the current terminology. The dataset (Pfeffer et al. 2023) used may contain personally identifiable information, and some posts may contain offensive content. Furthermore, there is a potential for biases arising from the platform's unique user demographics (Olteanu et al. 2019). Pfeffer et al. (2023) has been released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The data is shared within the boundaries of Twitter's terms of service. We obtained Institutional Review Board (IRB) approval from University of Southern California for our annotation experiment using Prolific annotators. There is the potential for misuse of the findings from our work; individuals might use causal language for manipulation, political campaigns, among others. Lastly, we note that it is possible that our causal language detection approach will fail for certain English-speaking linguistic subgroups.

Acknowledgments

We would like to express our gratitude to Jürgen Pfeffer for providing the dataset for this work. We would also like to thank Ashwin Rao, Bohan Jiang, Hunter Priniski, Kenny Joseph, Luca Luceri, Myrl Marmarelis, Siyi Guo, and Zhivar Sourati for their valuable comments on this work. This research was supported, in part, by MURI-ONR-N00014-20-S-F003 on Persuasion, Identity, and Morality in Social-Cyber Environments.

References

Adair, J. G. 1984. The Hawthorne effect: a reconsideration of the methodological artifact. *Journal of applied psychology*, 69(2): 334.

Aristotle. 350BCE. *The Metaphysics*. The Internet Classics Archive.

Back, M. D.; Stopfer, J. M.; Vazire, S.; Gaddis, S.; Schmukle, S. C.; Egloff, B.; and Gosling, S. D. 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychological science*, 21(3): 372–374.

Baumeister, R. F.; Bratslavsky, E.; Finkenauer, C.; and Vohs, K. D. 2001. Bad is stronger than good. *Review of general psychology*, 5(4): 323–370.

Berger, J.; and Milkman, K. L. 2012. What makes online content viral? *Journal of marketing research*, 49(2): 192–205.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

Brady, W. J.; Wills, J. A.; Jost, J. T.; Tucker, J. A.; and Van Bavel, J. J. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28): 7313–7318.

Cacioppo, J. T.; and Gardner, W. L. 1999. Emotion. *Annual review of psychology*, 50(1): 191–214.

Corrigan, R.; and Denton, P. 1996. Causal understanding as a developmental primitive. *Developmental review*, 16(2): 162–202.

Del Vicario, M.; Bessi, A.; Zollo, F.; Petroni, F.; Scala, A.; Caldarelli, G.; Stanley, H. E.; and Quattrociocchi, W. 2016. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3): 554–559.

Ekman, P. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60): 16.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Friggeri, A.; Adamic, L.; Eckles, D.; and Cheng, J. 2014. Rumor cascades. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 101–110.

Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N. F.; Peters, M. E.; Schmitz, M.; and Zettlemoyer, L. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 1–6.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Girju, R.; and Moldovan, D. I. 2002. Text Mining for Causal Relations. In *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, 360–364. AAAI Press. ISBN 157735141X.

González-Bailón, S.; Lazer, D.; Barberá, P.; Zhang, M.; Allcott, H.; Brown, T.; Crespo-Tenorio, A.; Freelon, D.;

Gentzkow, M.; Guess, A. M.; et al. 2023. Asymmetric ideological segregation in exposure to political news on Facebook. *Science*, 381(6656): 392–398.

Gopnik, A.; Schulz, L.; and Schulz, L. E. 2007. *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

Gopnik, A.; and Wellman, H. M. 2012. Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6): 1085.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, 55–130. Elsevier.

Guille, A.; Hacid, H.; Favre, C.; and Zighed, D. A. 2013. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2): 17–28.

Haber, N.; Smith, E. R.; Moscoe, E.; Andrews, K.; Audy, R.; Bell, W.; Brennan, A. T.; Breskin, A.; Kane, J. C.; Karra, M.; et al. 2018. Causal language and strength of inference in academic and media articles shared in social media (CLAIMS): A systematic review. *PloS one*, 13(5): e0196346.

Haidt, J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4): 814.

Hartmann, J. 2022. Emotion English DistilRoBERTa-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Heath, C.; Bell, C.; and Sternberg, E. 2001. Emotional selection in memes: the case of urban legends. *Journal of personality and social psychology*, 81(6): 1028.

Heider, F. 2013. *The psychology of interpersonal relations*. Psychology Press.

Henrich, J. 2001. Cultural transmission and the diffusion of innovations: Adoption dynamics indicate that biased cultural transmission is the predominate force in behavioral change. *American anthropologist*, 103(4): 992–1013.

Hofmann, W.; Wisneski, D. C.; Brandt, M. J.; and Skitka, L. J. 2014. Morality in everyday life. *Science*, 345(6202): 1340–1343.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960.

Hopp, F. R.; Fisher, J. T.; Cornell, D.; Huskey, R.; and Weber, R. 2021. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, 53: 232–246.

Jamieson, K. H.; and Cappella, J. N. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

Lazarus, R. S.; and Folkman, S. 1984. *Stress, appraisal, and coping*. Springer publishing company.

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; et al. 2009. Computational social science. *Science*, 323(5915): 721–723.

Li, M.; Wang, X.; Gao, K.; and Zhang, S. 2017. A survey on information diffusion in online social networks: Models and methods. *Information*, 8(4): 118.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marini, M. M.; and Singer, B. 1988. Causality in the social sciences. *Sociological methodology*, 18: 347–409.

Mercier, H.; and Sperber, D. 2017. *The enigma of reason*. Harvard University Press.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. 2011. Understanding the demographics of Twitter users. In *Proceedings of the international AAAI conference on web and social media*, volume 5, 554–557.

Murphy, G. L.; and Medin, D. L. 1985. The role of theories in conceptual coherence. *Psychological review*, 92(3): 289.

Norenzayan, A.; and Heine, S. J. 2005. Psychological universals: What are they and how can we know? *Psychological bulletin*, 131(5): 763.

Nyhan, B.; Settle, J.; Thorson, E.; Wojcieszak, M.; Barberá, P.; Chen, A. Y.; Allcott, H.; Brown, T.; Crespo-Tenorio, A.; Dimmery, D.; et al. 2023. Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, 1–8.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kıcıman, E. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2: 13.

Pearl, J. 2009. *Causality*. Cambridge University Press.

Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic books.

Pfeffer, J.; Matter, D.; Jaidka, K.; Varol, O.; Mashhadi, A.; Lasser, J.; Assenmacher, D.; Wu, S.; Yang, D.; Brantner, C.; et al. 2023. Just another day on Twitter: a complete 24 hours of Twitter data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 1073–1081.

Priniski, J.; Verma, I.; and Morstatter, F. 2023. Pipeline for modeling causal beliefs from natural language. In *Association for Computational Linguistics (Volume 3: System Demonstrations)*, 436–443.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rozin, P.; and Royzman, E. B. 2001. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, 5(4): 296–320.

Ruths, D.; and Pfeffer, J. 2014. Social media for large studies of behavior. *Science*, 346(6213): 1063–1064.

Sloman, S. 2005. *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Stieglitz, S.; and Dang-Xuan, L. 2012. Political communication and influence through microblogging–An empirical

analysis of sentiment in Twitter messages and retweet behavior. In *2012 45th Hawaii international conference on system sciences*, 3500–3509. IEEE.

Stieglitz, S.; and Dang-Xuan, L. 2013. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4): 217–248.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 1.

Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE second international conference on social computing*, 177–184. IEEE.

Thompson, S. C. 2016. Illusions of control. In *Cognitive illusions*, 134–149. Psychology Press.

Thompson, S. C.; Armstrong, W.; and Thomas, C. 1998. Illusions of control, underestimations, and accuracy: a control heuristic explanation. *Psychological bulletin*, 123(2): 143.

Traag, V. A.; Waltman, L.; and Van Eck, N. J. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1): 5233.

Trabasso, T.; and Van Den Broek, P. 1985. Causal thinking and the representation of narrative events. *Journal of memory and language*, 24(5): 612–630.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.

Webberley, W.; Allen, S.; and Whitaker, R. 2011. Retweeting: A study of message-forwarding in twitter. In *2011 Workshop on Mobile and Online Social Networks*, 13–18. IEEE.

Wolff, P.; Song, G.; and Driscoll, D. 2002. Models of causation and causal verbs. In *Papers from the 37th meeting of the Chicago Linguistics Society, Main session*, volume 1, 607–622. Chicago Linguistics Society Chicago.

Wright, D.; and Augenstein, I. 2021. Semi-Supervised Exaggeration Detection of Health Science Press Releases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10824–10836.

Xu, J.; Zuo, W.; Liang, S.; and Zuo, X. 2020. A review of dataset and labeling methods for causality extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1519–1531.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, please see the Limitation and Ethical Considerations.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, please see the Abstract and the Introduction.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, please see the paper.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, please see the Limitations and Ethical Considerations.
   (e) Did you describe the limitations of your work? Yes, please see the Limitations and Ethical Considerations.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, please see the Limitations and Ethical Considerations.
   (g) Did you discuss any potential misuse of your work? Yes, please see the Limitations and Ethical Considerations.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, please see the Limitations and Ethical Considerations.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, our paper conforms to the ethics review guidelines.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes, we cited them, please see the Dataset Description.

(b) Did you mention the license of the assets? Yes, please see the Limitations and Ethical Considerations. We did not mention the license for the models because they are publicly available.

(c) Did you include any new assets in the supplemental material or as a URL? No, we don't have any.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, please see the Limitations and Ethical Considerations.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, please see the Limitation and Ethical Considerations.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, please see the Validation and the Appendix.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, we obtained IRB approval from University of Southern California for our annotation experiment. Please see the Limitations and Ethical Considerations.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, please see the Appendix (Annotation).

(d) Did you discuss how data is stored, shared, and deidentified? Yes, we discussed how data was stored, shared, and deidentified. Please see the Limitations and Ethical Considerations.

Appendix

Annotation

We spent $200 in our annotation task, which included $50 in service fees to Prolific. Annotators, each assigned 30 messages to annotate via a Qualtrics survey link provided through Prolific, were paid $5 each, with an hourly wage of $30. Prolific automatically deidentified participant data.

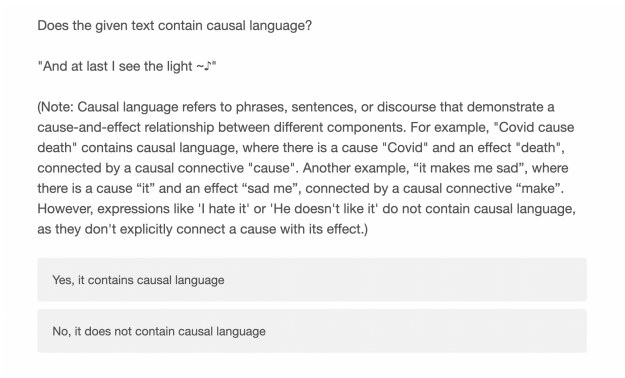


Figure 6: Annotation example

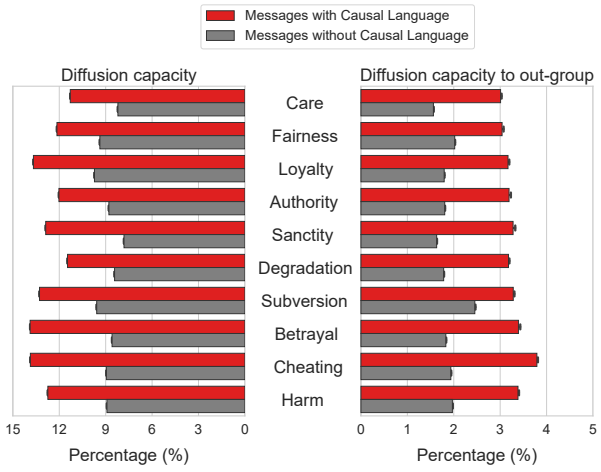
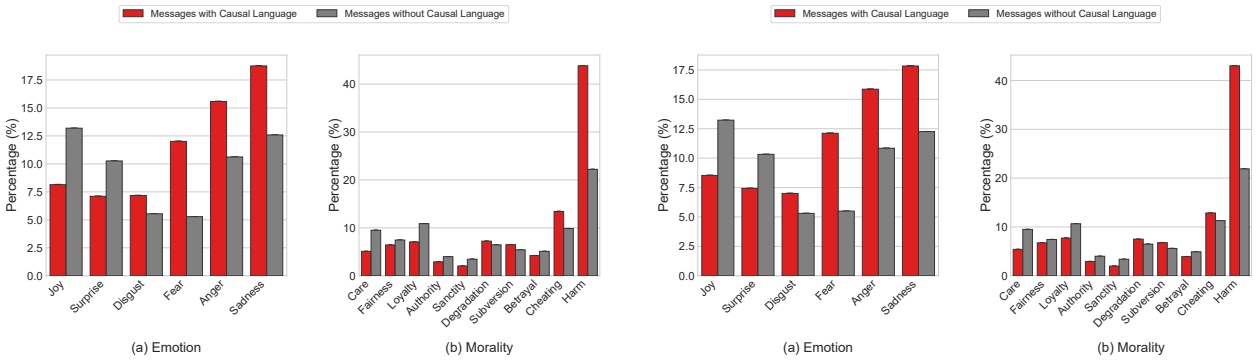
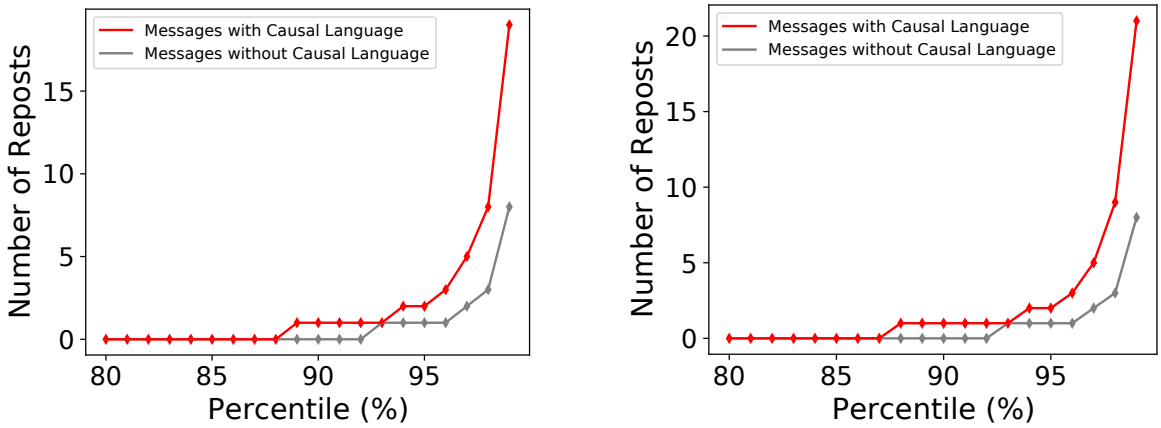


Figure 7: Differential dissemination patterns and out-group dissemination patterns of messages with or without causal language, across varied morality categories. Error bars represent variability across 100 runs with the Louvain algorithm.



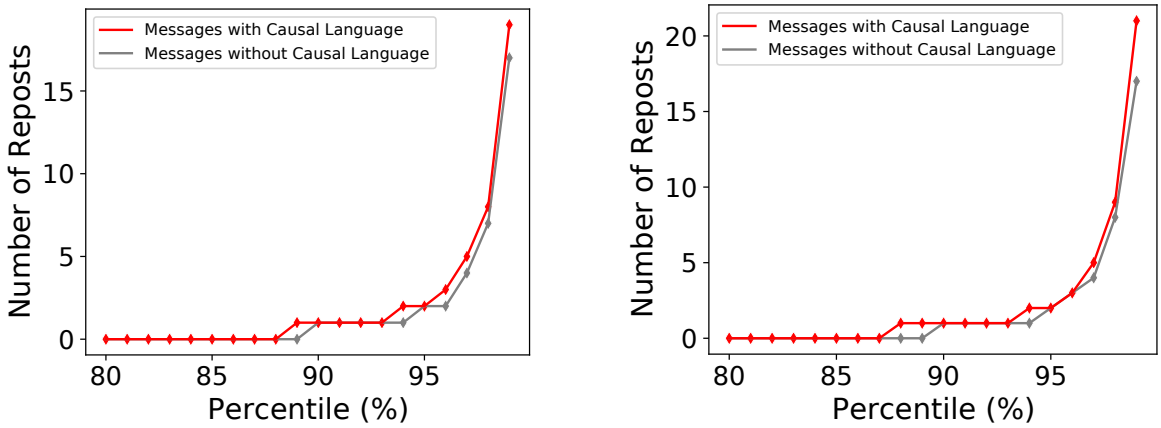
(a) Day 2 (b) Day 3

Figure 8: Comparative analysis of the percentage distribution of dominant (a) emotion or (b) morality categories present in messages on day 2 and day 3, segregated by the presence or absence of causal language. Error bars reflect the variability across 100 bootstrapping runs.



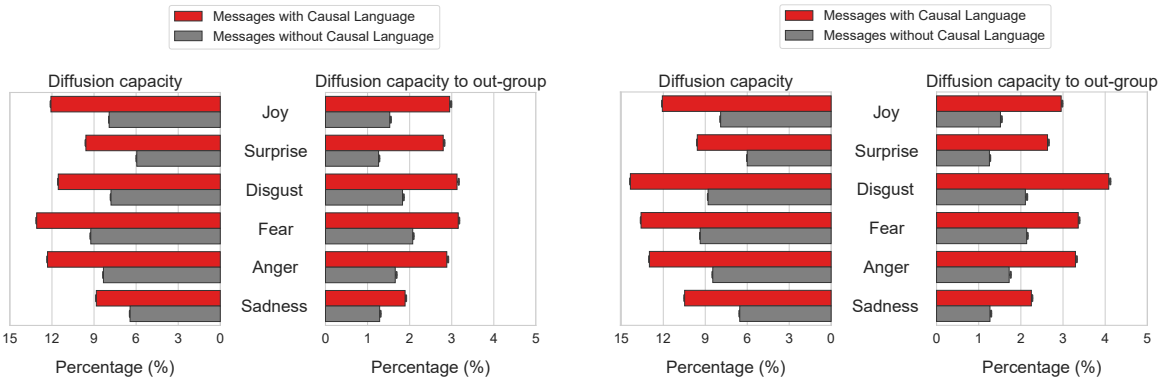
(a) Day 2 (b) Day 3

Figure 9: Repost distribution for messages employing causal versus non-causal language, on day 2 and day 3.



(a) Day 2 (b) Day 3

Figure 10: Repost distribution for paired messages employing causal versus non-causal language via counterfactual analysis, on day 2 and day 3.



(a) Day 2

(b) Day 3

Figure 11: Differential dissemination patterns and out-group dissemination patterns of messages with or without causal language on day 2 and day 3, across varied emotion categories. Error bars represent variability across 100 runs with the Louvain algorithm for community detection.