THE IMPACT OF CONTENT AND SENTIMENT COHERENCE ON INFORMATION DIFFUSION

CHEN ZIHAN

MPhil

THE HONG KONG POLYTECHNIC UNIVERSITY

2021

The Hong Kong Polytechnic University

Department of Management and Marketing

The Impact of Content and Sentiment Coherence on Information Diffusion

CHEN Zihan

A thesis submitted in partial fulfilment of requirements for the degree of Master of Philosophy

May 2021

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

CHEN Zihan

Abstract

While online discussion has been extensively studied in the previous literature, the role of information coherence along the course of information diffusion remains underexplored. In this study, I investigate the influence of information coherence in the online discussion context with a focus on the sequential patterns of user posts. I measure the information coherence using text mining techniques by two dimensions: content and sentiment. Using the data from popular online automobile discussion platforms in China, I empirically estimate the effect of information coherence on the duration and breadth of information diffusion. The empirical evidence sheds light on the heterogeneity of the coherence influence. It indicates that relevant content motivates more replies in a shorter duration, whereas consistent sentiment is associated with longer duration and fewer replies. The moderating analysis on the topic further deepens the understanding of the relationship between information coherence and information diffusion. I found that under the topic of information acquisition (vs. experience sharing), relevant content and consistent sentiment can end the discussion about information acquisition faster with fewer replies. Besides, the moderating analyses on user interaction suggest that user interaction does not always increase participation in terms of more replies and longer duration. High content dependency (vs. low content dependency) increase participation more efficiently under low content coherence, whereas high original poster’s replies (vs. low original poster’s replies) increase participation more efficiently under high sentiment coherence. The findings provide managerial insights for platforms to facilitate online discussion and for companies to facilitate product management.

Acknowledgements

I would like to thank my supervisors, Dr. FENG and Prof. NGAI, for the valuable advice and support they have given me in the writing of this thesis.

Table of Contents

List of Tables

List of Figures

The Impact of Content and Sentiment Coherence on Information Diffusion

1. Introduction

Online discussion has become prominently prevalent in our daily life and facilitates information exchange among online users. Due to the ongoing COVID-19 pandemic and social distancing requirement, more and more people start using the internet to communicate with others and spend more time online to facilitate their learning and working process. According to a recent report , online users have grown more than 10 percent, which lead to more than half of the world population have participated in the online discussion as of July 2020. Students are more actively engage in online discussion forums as a way to discuss class-related questions (Hill and Fitzgerald 2020), and people participate in online forums to share and search for useful information and personal experience (Kornfield and Toma 2020, Rawaf et al. 2020). Both firms and customers can benefit from the huge volume of user interactions via information diffusion in online discussion (Manchanda et al. 2015; Wang and Chaudhry 2018).

Different from the contexts of social media where users join in discussions with a key motivation to build social connections, online discussion platforms are more dedicated to the function of information exchange with a unique feature of displaying user posts with the sequence. In other words, the sequential patterns of user posts are highlighted such that the latter users will read the former information before they reply. Therefore, user's behavior in information diffusion is easily affected by other users' opinions and managerial responds they saw previously (Godes and Silva 2012, Wang and Chaudhry 2018, Wang et al. 2018). Despite the recent attention on the sequence in the online discussion, how the inherent information pattern affects information diffusion among acquaintances remains unclear. Few case studies have suggested that the content and form of information transform during the diffusion process (Im et al. 2011, Kane et al. 2014), which sheds light on the importance of sequence as well as the need to explore the information inside messages. To bridge the gap, I focus on the role of information coherence in online discussion, and examine its influence on the duration and breadth of information diffusion.

How online users create and share information is an important topic in informatics literature. Coherence in a group along with its sequential structure is regarded as the reason why communication influences its group output (Pavitt and Johnson 1999). Linguistics defines coherence as “a continuity of sense”, being the “mutual access and relevance within a configuration of concepts and relations” (De Beaugrande and Dressler 1981). In offline discussion, coherence is viewed as utterance relevant among speak turns across people and measured by the conditional probability of topic change each turn (Pavitt and Johnson 1999). In online discussion, coherence is often regarded as the coherence relationship among texts and measured by the reply-to relationship between messages (Barzilay and Elhadad 1999, Fu et al. 2008). However, despite the multi-dimensional nature of coherence, researchers often focus on one type of coherence yet ignore the others (Givón 1995, Korolija 2000, Picciotto 2005). Therefore, focusing on the thread-level information diffusion in the online discussion, this research measures information coherence in online discussion by two dimensions: content and sentiment, and defines information coherence as the content relevance and sentiment consistency conveyed over the sequence of user posts within a thread. This separation view of information is inspired by the lab experiments on flagging fake news, where researchers do not distinguish between content and sentiment, yet their results shed light on the difference between them (Xiao and Benbasat 2015, Ananthakrishnan et al. 2020). Although sentiment in the flagged fake news is insignificant on changes user’s belief (Ananthakrishnan et al. 2020), content design on identifying bias increases users’ ability to correctly detect fake news (Xiao and Benbasat 2015). Although the existing studies on fake news diffusion cannot represent the inherent pattern in online discussion, it does imply the difference between sentiment and content. In this study, I emphasize on the heterogeneity effect of information coherence in the information diffusion process in the online discussion context.

How, then, does coherence influence discussion? There are no consensus findings in the literature. Some studies find that people prefer opinion-consistent information (Ehrlich et al. 1957, Streufert 1973, Villarroel Ordenes et al. 2017), while others argue that people tend to participate in controversial discussions (Festinger 1957, Freedman 1965, Chen and Berger 2013). One explanation is that the controversial findings may result from the often-one-dimensional definition of coherence, as people tend to sort and define things on the basis of a single dimension (Medin et al. 1987). An alternative explanation of these controversial findings is that, the difference in coherence influences may result from a variety of context-dependent features (Festinger 1957, Sears and Freedman 1967, Chen and Berger 2013), ranging from individual difference such as the knowledge of subjects (McNamara and Kintsch 1996) to usefulness (Freedman 1965) and interestingness (Chen and Berger 2013) of the information being discussed. In my study, I consider two groups of potential moderators – discussion topic and user interaction, which might moderate the relationship of information coherence and information diffusion. First, in the online discussion, topics can significantly change people's participation behavior (Thomson 2006, Chen and Berger 2013). Even in less structured discussions like online chats, topics may influence coherence patterns throughout the discussion (Stromer-Galley and Martinson 2009). Second, one of the unique features of an online forum is that user interacts with each other, as the latter users will read the former information before participating in online discussion. The existence of such a sequential pattern has been linked to the dependency among users, as opinions that formed during the previous discussion have a direct impact on the latter opinions in online discussion (Dolatabadi et al. 2020). And indeed, research has shown that dependency can mitigate the negative consequences caused by incoherence (Wang et al. 2010). Besides, original poster replies can indicate the reciprocity effect of user interaction. Users often take different roles in online discussion and affect the direction of discussion differently (Chan et al. 2010, Hecking et al. 2016). Yet reciprocity is a reason why users participate in online discussion (Lakhani and Von Hippel 2004). In the asynchronous discussions of an online course, students often feel more satisfied about the course with the contact and feedback from instructors (Swan 2002). In a more general online discussion, the effect of reciprocity also exists (Faraj and Johnson 2011, Johnson et al. 2014), and I believe that when original posters who set the topic lead the discussion and interact with other users, it may moderate the effect of information coherence. Even though previous literature has provided pieces of information on the effect of information coherence, there has been limited investigation of the possible interactions while taking into account heterogeneity in information coherence. A more systematic investigation is needed to facilitate our understanding of the information coherence (i.e., whether and how the controversial findings result from heterogeneity in information coherence and potential moderating effects respectively) and how the platform administrators could take advantage of its effect to approach their goal (e.g., facilitate user engagement). In this research, I further explore the heterogeneity of information coherence by exploring the moderating effect of topic and user interaction.

Further, I examine whether the effect of information coherence is subject to increasing or diminishing returns on information diffusion. I posit that the direction of this effect (i.e., square terms of information coherence) depends on a central trade-off. On the one hand, the theory of cognitive dissonance and the public nature of thread discussion suggests that users tend to avoid participating in dissonance online discussion (Festinger 1957). In other words, the information diffusion should spread (i.e., more replies and longer duration) with the marginal increases of information coherence. On the other hand, if discussing the information is the primary motive for information diffusion, cognitive load theory indicates that processing relevant content and consistent sentiment might impose information overload and cause a feeling of redundancy (Sweller 2011). As such, information diffusion should spread less as the marginal effect of information coherence increases.

In terms of the consequence of information coherence, I study information diffusion based on duration and breadth. The duration of discussion is quantified by the count of days that the thread has received at least one reply. Breadth refers to replies in the thread (i.e., number of replies equals the maximum length of diffusion chain), resulting from the single-chain feature in the online discussion forum. I focus on these two aspects for the following reasons. First, it enables us to explore the trade-off between duration and replies raised by the effect of information coherence. To the best of my knowledge, the present work is the first to explore the tension between duration and replies in the online discussion (Hoffman and Schraw 2009, Hoffman 2010, Lehmann et al. 2012, Ibrahim et al. 2017, Jaakonmäki et al. 2017). Second, empirical evidence has linked changes in the coherence to changes in diffusion (Streufert 1973, Herring 1999, Weger and Aakhus 2003, Villarroel Ordenes et al. 2017). However, there is still limited research exploring the heterogeneity of information coherence on information diffusion using empirical data. Besides, understanding the heterogeneity of information coherence helps platform administrators to modify their platform's target. Thus, more formally, I seek to answer the following research questions:

Q1: How does heterogeneity in information coherence, represented by content coherence and sentiment coherence, influence information diffusion, in terms of duration and number of replies?

Q2: How does discussion topic and user interaction moderate the relationship between information coherence and information diffusion, respectively?

To answer these questions, I conduct my empirical analyses using the data from leading online automobile discussion platforms in China. I first use text mining methods to derive the information coherence and dependency measures and to infer the topics. As information diffusion might also affect coherent degree, I also construct instrumental variables to identify the effect of information coherence. The instrumental variables hinge on my ability to observe the information coherence of the adjacent previous threads that the users in the focal thread participated in. By examining the effect of information coherence through instrumental variables, I can establish causality in my models. In the online discussion in which users are mostly acquaintances with common interests, information coherence can serve as a signal of information relevance and consistency to draw in more participants (Aral et al. 2009, De Choudhury et al. 2010, Mason et al. 2011, Zare et al. 2020).

I apply the survival analysis (i.e., semi-parametric Cox-proportional modelling) to evaluate the impact of information coherence on duration, and use Poisson regression with endogenous regressors (i.e., Poisson GMM) to identify the effect of information coherence on number of replies. The results show the heterogeneous and non-linear effect of information coherence. It indicates that higher content coherence increases replies in shorter duration, whereas higher sentiment coherence leads to the opposite (fewer replies in longer duration). There is a curvilinear relationship between sentiment coherence and information diffusion, such that users tend to discuss more and longer (more replies and longer duration) at a moderate level of sentiment coherence. I also explore the moderating effects of the topic and user interaction further to understand the reasons behind the role of information coherence. Comparing to the discussion topic of experience sharing, the discussion of information acquisition has fewer replies with a shorter duration. It suggests with higher information coherence, users tend to end the thread quickly with fewer replies in the discussion about information acquisition (vs. experience sharing) when the previous consensus has already been achieved, possibly due to the increasing feeling of redundancy. In addition, I measure user interaction by thread dependency and original poster's reply and evaluate their moderating effects. The results suggest that user interaction cannot always lead to more participation in terms of duration and replies.

This research has several contributions to the existing literature. First, I extend the literature on the antecedent of online discussion. My work is built on recent research of re-evaluating the magnitude of coherence in online discussion. Abbasi et al. (2018) argue that mainly focusing on reply-to relations while ignoring interactions among texts might result in difficulties in accurately revealing the coherence in online discussion. They find that when including such features through a language-action perspective-based framework, their predictions of discussion patterns can outperform existing methods that are devoid of conversation structure information. Stromer-Galley and Martinson (2009) conduct a comparison study in online chat rooms to investigate coherence when people synchronously discuss online. They find evidence of a relatively high coherence among political discussion, which contradicts previous findings that synchronous online discussion suffers from low coherence (Weger and Aakhus 2003). In this study, I investigate the influence of information coherence on information diffusion, emphasizing its role as an important antecedence of online discussion. The results show that the heterogeneous and nonlinear effects of information coherence do exist, and it affects the duration and breadth of information diffusion. The marginal effect of content coherence is driven by consonance, whereas the marginal effect of sentiment coherence is driven by redundancy. Besides, I explore two groups of potential moderating factors on information coherence – topic and user interaction and delve into how consonance and redundancy drive information coherence. The results show that the effect of information coherence ends the discussion faster in fewer replies when the topic is information acquisition, indicating the feeling of redundancy more likely to happen under the topic of information acquisition. The results on user interaction suggest that the effect of information coherence not always increases with more user interaction. The moderating effect of dependency increases diffusion (i.e., longer duration and more replies) more under low content coherence, whereas the moderating effect of the original poster's replies increases diffusion more when sentiment coherence is high in the thread. These results further support the heterogeneity effect of information coherence on information diffusion. Last, I apply a data forecasting method to investigate the dependency among user interactions in a thread. The dependency is represented by the inversed smoothing factor from the exponential smoothing model. It offers a sophisticated way to quantify the extent to which the similarities of the latter posts rely on the similarity patterns of earlier posts in the thread. Unlike the previous research that focuses on pairwise user interaction in social networks (Viswanath et al. 2009, Wang et al. 2018), my dependency measurement captures interdependent patterns among thread-level user posts.

The rest of the paper is organized as follows: In Section 2, I briefly summarize the literature related to online discussion, information coherence, and information diffusion. In Section 3, I discuss my research setting and variables of interest. In Section 4, I propose my model and identification strategy to casually evaluate the effect of information coherence and discuss the main results. Section 5 discusses the model extensions and robustness checks. Section 6 concludes the paper.

2. Related Literature

2.1. Dynamics of Online Discussion

Online discussion denotes that users exchange opinions, feelings, or experiences on a topic in online communities. A stream of literature has focused on the generation of online discussion (or user-generated content: UGC) through different perspectives and emphasized the dynamics of online discussion such that subsequent user posts are influenced by previous information (Moe and Trusov 2011, Wang et al. 2018). From a social influence perspective, Wang et al. (2018) investigate how friends tie influence online ratings. They find that online users are socially nudged and tend to follow friends' previous patterns. Godes and Silva (2012) explicitly point out that previous research has neglected the inter-correlations across online reviews. However, users on pure review platforms are less likely to interact with each other, as their reviews can be written based on their own product experience in many cases. In contrast, reading previous posts by sequence is an inevitable endeavor and thus becomes the unique feature of participating in the online discussion forums. Recent papers have studied the effect of firm interventions on subsequent opinions. For example, Wang and Chaudhry (2018) study how managers' responses to previous users' reviews affect subsequent reviews. They calculate the sentiment of users' reviews and the similarity of managers' responses to the reviews. Their findings suggest that managers' responses to the negative reviews help to fix problems, thus leading to a positive impact on subsequent opinions. However, those responses to the positive reviews are treated as deliberate promotion and result in a lower subsequent opinion. While information similarity in the review threads has been underlined, their analysis is from the firms' standpoint and applied to the platforms which encourage firms to respond to customers' feedback. Ananthakrishnan et al. (2020) study the subsequent changes in user behavior after the platform displays flagging fake news together with true news in the lab experiment on restaurant reviews. Their findings suggest that users prefer to consume in a restaurant with a lower historical record of fake news after viewing the flagging news. While the fake information may exhibit discontinuity from the true one in the discussion series, a fundamental question remains unaddressed such that how the inherent information pattern along the course of discussion affects information diffusion. In this research, I focus on understanding the sequential pattern of user posts over time in the online discussion context.

Despite the rich literature on online discussion, few have examined the change of information pattern through the information diffusion process. One exception is the work of Im et al. (2011). They study the evolvement of news via case studies in online discussion and find that the content and form of news transformed across the diffusion process. Their analysis emphasizes the importance of information content as well as information sequence, as the latter version of news might change according to the information in the previous version of the news. In this research, I aim to investigate the effect of the inherent information pattern (i.e., coherence across user posts over time) on information diffusion in the online discussion context, where users are mostly acquaintances with common interests and are self-motivated to participate in the discussion.

2.2. Information Coherence

Previous studies have stated the lack of coherence in user discussion (Herring 1999, Weger and Aakhus 2003, Honey and Herring 2009). For example, Weger and Aakhus (2003) study argumentation in online chat rooms, and they find that there is a lack of coherence and this incoherence discourages user participation in the chat rooms. Herring and Nix (1997) indicate that the lack of coherence is a result of simultaneous turn-taking design in the chat rooms. In the words of Herring (1999), text-based online discussion is "interactionally incoherent due to limitations imposed by messaging systems on turn-taking and reference, yet its popularity continues to grow". In order to make their message more readable, online users might try to increase the coherence in the discussion thread through various strategies, such as providing contextual information and lexical repetition (Te'eni 2006, Woerner et al. 2007).

Coherence has been viewed as a multi-dimensional construct (Korolija 2000) and with multiple features (Bou-Franch et al. 2012, Abbasi et al. 2018). Following the literature, I incorporate two dimensions of information coherence: content and sentiment. The first dimension is content coherence. Streufert (1973) conduct laboratory experiments and show that receiving the relevant information to some events can positively influence the volume of group decisions making under complex environments. Cheung et al. (2008) find that the content of electronic word-of-mouth is a key factor in adopting online opinions. In this study, I use content coherence to indicate the extent to which the content among user posts in the online discussion is relevant to each other. The second dimension is sentiment coherence. Villarroel Ordenes et al. (2017) investigate the sentiment incoherence across sentences within text-based reviews. They find that an increase of sentiment incoherence has a negative effect on the overall sentiment strength, because the sentiment expresses a set of sequentially organized propositions to explain an overall opinion and the use of contradictory sentiment expressions might convey a less degree of conviction. A similar point has been raised by Das and Chen (2007) that studies cannot disregard patterns of sentiment across sentences. In this study, I use sentiment coherence to indicate the extent to which the sentiment among user posts in the online discussion is consistent with each other. However, previous researchers did not distinguish sentiment from content. They either regard sentiment coherence and content coherence as in one feature (Deng et al. 2011, Chen and Berger 2013, Abbasi et al. 2018) or only consider one type of information coherence (Villarroel Ordenes et al. 2017). Evidence from lab experiments on fake news supports the heterogeneity effect of information coherence. Ananthakrishnan et al. (2020) randomly assign subjects to expose to the negative, positive, or mixed sentiments of fake news. They do not observe changes in subjects' beliefs across the different sentiments of fake news. While Xiao and Benbasat (2015) find that content design on fake news warning, such as adding advice on risk handling design and framing of the advice, affect subjects' ability to detect bias in the online recommendation. These results imply the difference between sentiment and content in information.

In this study, I consider information coherence based on the sequence of user posts and investigate the information coherence in the online discussion using the linguistic features of user posts. I measure the information coherence in terms of content coherence and sentiment coherence, as the proxies of content relevance and sentiment consistency in the online discussion. Particularly, I am interested to explore the heterogeneity effect of information coherence on information diffusion.

2.3. Information Diffusion

Previous researchers define information diffusion as the process that information spreads among entities, in a close environment (Wan and Yang 2007, Guille and Hacid 2012). It has been widely studied in a variety of online environments ranging from emails (Aral et al. 2007, Iribarren and Moro 2009) to the online discussion (De Choudhury et al. 2010, Li et al. 2017). In the online discussion, information diffusion has been measured in terms of quantity and time duration (Yang and Counts 2010, Zhang and Peng 2015). In this study, I define information diffusion as information spreads among users within the thread. I consider duration and breadth of information diffusion in the online discussion forum. The duration reflects the number of active days that a thread has received user posts, whereas the breadth of information diffusion refers to the total number of replies in the thread.

2.4. Information Coherence and Information Diffusion

The theoretical tension on the effect of information coherence on information diffusion has been documented by different researchers. Previous literature obtain controversial findings on the effect of information coherence on the breadth of diffusion (i.e., replies). Weger and Aakhus (2003) explore features in online chat rooms, and they find the online conversation is led by incoherence. They explain this result from the simultaneous user interaction design. Users' replies are highly likely to be ignored by other users due to the simultaneous interruption, which might demotivating users to join the discussion. Villarroel Ordenes et al. (2017) investigate the sentiment incoherence across sentences within text-based reviews. They find that an increase in sentiment incoherence leads to a decrease in overall sentiment strength. It is because that the sentiment expresses a set of sequentially organized propositions to explain an overall opinion and the use of contradictory sentiment expressions might convey a less degree of conviction. However, when studying coherence in the online discussion, Herring (1999) finds that loosened coherence might increase user interaction as a result of language play. This language play makes participating in the online discussion become a more attractive thing to do for users. Besides, as mentioned in the work of Stromer-Galley and Martinson (2009), a coherent discussion pattern might exist among an active group of users. Using data from online news websites and laboratory experiments, Chen and Berger (2013) indicate that more discussions are correlated with the articles that include moderately controversial opinions.

However, to the best of my knowledge, few researchers investigate the effect of information coherence on the duration of information diffusion. In an online discussion, a reply may be created on a different day from the previous discussion it responds to (Herring 1999). Stromer-Galley and Martinson (2009) study coherence in online discussions in topic-oriented chat rooms. They admit their limitation on not considering discussion across days or in longer duration and argue that including such data into consideration will increase the robustness of their findings of coherent discussion across topics. Thus, in this study, I consider the heterogeneous effect of information coherence on two aspects of information diffusion, namely, duration and breadth of information diffusion. The moderating effect of topic and user interaction further uncover the heterogeneous effect of information coherence.

In addition, previous research shows the importance of cognitive processes in driving information diffusion in online discussion. Cognitive dissonance theory indicates that people are not in favor of cognitive dissonance, resulting in their avoidance of cognitive inconsistency (i.e., favor of cognitive consistency) while processing messages (Festinger 1957). For example, Moreland (1987) study the formation of groups and find that coherence is expected to see in group formation, especially when groups are formed within people who share similar interests. Alexander (1964) examines drinking behavior in friend groups. His study suggests that a coherence group where all people are drinkers is perceived to be more attractive and popular than an incoherence group where only a proportion of people are drinkers. In an online discussion, higher information coherence can be perceived as having a consonance discussion, which attracts more replies in a longer duration. Therefore, if cognitive dissonance drives information diffusion under high information coherence, I expect that the marginal effect of information coherence positively influences information diffusion in terms of duration and replies. On the contrary, cognitive load theory indicates that when processing unnecessary information, the huge amount of cognitive load might result in the feeling of redundancy (Sweller 2011). Working memory has the limitation on capacity and duration (Miller 1956, Peterson & Peterson 1959). When the previous discussion already reached high content relevance and sentiment consistency, users might have the feeling of redundancy and be unwilling to further diffuse the information to save their working memory. Therefore, if the cognitive load drives information diffusion under high information coherence, I expect that the marginal effect of information coherence negatively influences information diffusion in terms of duration and replies. In this study, I examine the nonlinear relationships of information coherence on information diffusion in the main analysis and explore the marginal effect of information coherence.

3. Research Methods, Text Mining, and Variables

3.1. Research Setting and Data

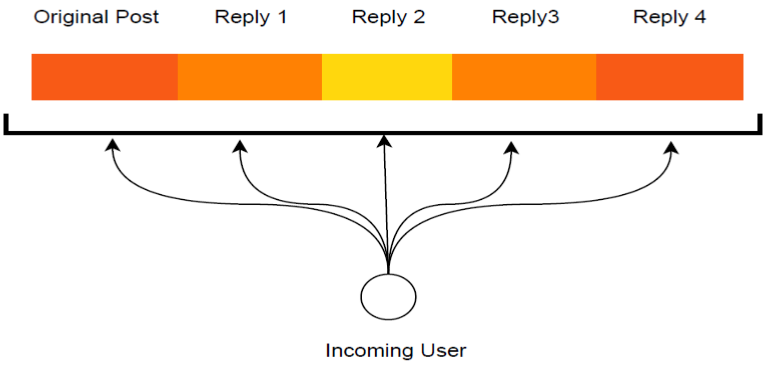
I use the data from two leading online automobile forums in China. Both platforms have more than 3 million daily active users on average, who engage in the online discussion by posting their opinions, experiences, and feelings related to automobile topics. The collection of posts is called thread. A thread is created by an original post (could be a question or description based on self-experience) and followed by users' (including the initiator's) replies under this thread. The replies are displayed along time sequence.

The dataset consists of 31,507 threads related to one automobile brand from the platform between January 2018 and October 2018. Among them, 2,421 threads have fewer than five posts, which are not long enough to detect sequential patterns and are removed from the final sample. I also remove the threads that contain the posts deleted by the administrators of platforms, indicating misinformation provided in those threads. Therefore, the final sample ensures an overall informative and rational online discussion setting and reveals sequential patterns in user posts. I also remove 72 threads with the users who do not have any historical data, hindering us from constructing the instrument variables. The final sample consists of 73,444 users and approximately 1.4 million posts from 29,014 threads. On average, each thread has 47 replies, ranging from 4 replies to 3,384 replies. The average active comment duration is 5 days (i.e., count of the dates that receive at least one reply), with a variation from 1 day to 98 days. The likelihood-ratio test suggests that there is a significant overdispersion on replies as well as duration ($p<0.01$).

3.2. Content Coherence and Sentiment Coherence

The key independent variable, information coherence, is measured by two dimensions, i.e., content and sentiment. To measure content coherence (*CONC*), I apply the Jaccard coefficient to calculate the lexical similarity between each pair of adjacent posts along the sequence of user posts, and aggregate the similarities to the thread level after taking a softmax weight. The softmax weight is a commonly used method in neural networks by computer scientists (Gao and Pavel 2017, Peng et al. 2017, Wang et al. 2018). It helps us convert numeric weights to a probability distribution that ranges from 0 to 1. To further illustrate, suppose that there are *n* replies (i.e., *n+1* posts) in a thread with index *j*. The weight on the similarity between i^{th} reply (i.e., $(i+1)^{th}$ post) and its previous post in thread *j* follows an exponential decay, and I normalize it by value sum of all the exponential decay in thread *j* to build the softmax weight, i.e., $w_{j,i} = \frac{e^{(-\frac{i-1}{n-1})}{}_j}{\sum_{i=1}^{n} e^{(-\frac{i-1}{n-1})}{}_j}$, with *i=1,2,..., n*. The logic is as follows: I assume that for an incoming user, she is likely to give most of her attention to the original post as well as the last reply she could observe in the thread (Wise et al. 2012). Her attention gradually decays when she moves from two edges to the middle of the replies in the thread. Figure 1 illustrates such attention distribution. I further assume that all the users in this thread following the same pattern. When summing up attention from all the users who participate in a thread discussion (e.g., thread *j*), the attention distribution is quite similar to the distribution of exponential decay $e^{(-\frac{i-1}{n-1})}{}_j$, with the original post receives the most attention from the participants, and the attention decreases with the sequence of replies. It has been shown that the exponential decay pattern exists in the online environment, such as in online citations and online video evolution (Avramova et al. 2009, Della Briotta Parolo et al. 2015). To adjust for the potential heavy tail popularity decay that is sometimes found in the online environment (Avramova et al. 2009), I first form the exponential decay before normalization when the weight of the last reply is not too small (i.e., 0.37). Then I normalize the result and transform it into a softmax weight.

Figure 1. Incoming User's Attention Distribution



Notes. Figure 1 shows how an incoming user's (i.e., who will generate the 5th reply) attention distributes across posts. The colour of the bar indicates the amount of attention this focal user puts on each post, the darker the colour, the more attention from this participant. I argue that the user put most of the attention on the original post and 4th reply, whereas put little attention on the middle reply.

Under this situation, content coherence of thread *j* can be generalized as a softmax weighted similarity measurement, which can be calculated as:

$$CONC_j = \sum_{i=1}^{n} w_{j,i} \times Jaccard(A_{j,i}, A_{j,i-1}) \tag{1}$$

where $Jaccard(A_{j,i}, A_{j,i-1})$ is the Jaccard similarity between *i*th reply and *(i-1)*th reply. $A_{j,i}$ is the word vector of the *i*th reply in thread *j*. Note that, in particular, $A_{j,0}$ represent the word vector of the original post in the thread *j*. In an online discussion, the reply is often short compared to articles and with repetition use of words. Previous research has shown that the Jaccard similarity coefficient has better performance for analyzing content from small and diversify documents (Deng et al. 2012, Tumuluru et al. 2012). Besides, the Jaccard similarity does not reduce when there is repetition use of words within a reply (AbuSafiya 2020).

To measure sentiment coherence (*SENC*), I use the Jieba toolkit for word segmentation. I adopt the sentiment dictionaries in Chinese from the Linguistic Inquiry Word Count (LIWC) program (Pennebaker et al. 2007). I retrieve the positive and negative words of each reply and original post and construct a vector consisting of three components: percentage of positive, negative, and neutral words. Sentiment coherence is measured by exponentially weighting the Cosine similarities between each pair of adjacent replies based on the sequence of user replies in each thread. Previous studies suggest that using cosine similarity to classify sentiment result in higher classification accuracy (Bhattacharjee et al. 2015, Thongtan and Phienthrakul 2019). As LIWC retrieve the percentage of positive, negative, and natural words from Chinese words in replies, it enabled me to create vectors with equal dimensions and calculate sentiment Cosine similarity between posts. Similarly, sentiment coherence in thread *j* can be calculated as:

$$SENC_j = \sum_{i=1}^{n} w_{j,i} \times Cosine(B_{j,i}, B_{j,i-1}) \tag{2}$$

where $Cosine(B_{j,i}, B_{j,i-1})$ is the Cosine similarity between *i*th reply and *(i-1)*th reply. $B_{j,i}$ is the three dimensions sentiment vector of the *i*th reply in thread *j*. Similarly, $B_{j,0}$ represent the sentiment vector of the original post in the thread *j*. Both coherence measures take values between 0 and 1, and a higher value indicates a greater degree of information coherence in the thread.

The descriptions of the main variables used in my empirical analyses are presented in Table 1. Specifically, the average of *CONC* over threads is 0.11, which is lower than the mean of sentiment coherence (i.e., 0.92, see the result in Table 2). It suggests that the sentiment is overwhelmingly coherent while the content is a bit diversified in the threads in my research context. This is probably because users tend to show emotional support by following previous posts but provide different content to add new information. Another reason could be that I have included the neutral dimension in the calculation of sentiment coherence. The high level of sentiment coherence might result from people's preference to use neutral words in automobile discussion contexts.

Table 1. Description of Main Variables

Variable	Description
$Duration_j$	Count of active days when there is at least one reply in thread *j*.
$Replies_j$	Number of replies in thread *j*.
$CONC_j$	Content coherence of thread *j* based on weighted Jaccard similarity.
$SENC_j$	Sentiment coherence of thread *j* based on weighted Cosine similarity.
$Topic_j$	The topic of thread *j*. It is a binary variable where 1 indicates the topic of information acquisition, and 0 indicates the topic of experience sharing
$COND_j$	Content dependency of thread *j* based on the series of Jaccard similarity values.
$SEND_j$	Sentiment dependency of thread *j* based on the series of Cosine similarity values.
$Original_j$	Log-transformed number of replies generated by the original poster in thread *j*.
$AveLen_j$	Log-transformed average reply length in thread *j*.
$OriLen_j$	Log-transformed length of the original post in thread *j*.

Table 2. Descriptive Statistics and Correlations (N =29,014)

	Variables	Mean	SD	Correlation									
				1	2	3	4	5	6	7	8	9	10
1	$Duration_j$	4.59	6.47	1									
2	$Replies\ _j$	46.6	60.36	0.51	1								
3	$CONC_j$	0.11	0.04	-0.09	-0.03	1							
4	$SENC_j$	0.92	0.1	-0.1	-0.45	0.09	1						
5	$Topic_j$	0.73	0.44	-0.07	-0.33	0.09	0.35	1					
6	$COND_j$	0.86	0.31	0.12	0.18	-0.05	-0.1	-0.09	1				
7	$SEND_j$	0.76	0.37	0.07	0.17	0	-0.15	-0.11	0.11	1			
8	$lnOriginal_j$	1.71	1.44	0.23	0.67	-0.02	-0.55	-0.39	0.17	0.17	1		
9	$lnAveLen_j$	2.68	0.64	-0.13	-0.48	0.07	0.59	0.43	-0.13	-0.17	-0.55	1	

10	$lnOriLen_j$	4.9	1.67	0.2	0.54	-0.12	-0.52	-0.43	0.13	0.14	0.54	-0.56	1

3.3. Contingent Factors

I explore the potential heterogeneity in the effect of information coherence by analyzing the moderating effects of two groups of contingent factors, namely, discussion topic and user interactions. The corresponding analyses and results are shown in section 5.

3.3.1. Topic

Previous studies have shown that discussion topic might influence a user's participation behaviour (Chaiken 1980, Thomson 2006, Stromer-Galley and Martinson 2009, Blau and Barak 2012, Chen and Berger 2013). I infer the topic of each thread based on its original post. Previous research has shown that the first post reflects the intention of the discussion and attracts users who share similar interests to participate (Liu et al. 2018).

I implement the latent Dirichlet allocation (LDA) (Blei et al. 2003) method to generate the topics. I pre-process the original posts of the threads with the Jieba toolkit for word segmentation and stop word removal. The resulting corpus of the original posts is transformed into a vector space of 147,068 words that comprise my dictionary. LDA is an unsupervised model that infers the topic of each thread by returning an underlying set of topic probabilities. I explore the number of underlying topics by model validation ranging from 2 to 15 topics. I also examine the perplexities with larger topics by testing the model results with 30, 50, and 100 topics. The model with 2 underlying topics is finally chosen because it produces the lowest perplexity, as shown in Table 3. One possible reason for the better performance of smaller topics is that I focus on automobile forums. Unlike forums having extensive discussions (e.g., Quora), the automobile forums only discuss topics related to the automobile. Besides, I only focus on a product from one brand, which further narrows the discussion scope. I assume that each thread only has one topic following the previous work related to thread discussion (Shen et al. 2006, Bhatia and Mitra 2010). Based on the topic probabilities generated by the LDA model, I assign the topic with a higher probability to be the topic of the thread. Table 4 summarizes the most frequently used words belonging to each topic I infer from LDA and the examples of original posts that I have translated from Chinese.

Table 3. Perplexity of LDA Across Topics

#Topic	**2**-15	30	50	100
Perplexity	10911	6,924,413	1,172,425,333	1.285e+17

Table 4. Keywords and Examples of Each Topic

Topic	Key words	Examples
Topic 1: Information acquisition	Oil Consumption, Automobile Brand, Space, Problem, Function, Compare	1. I want to ask those who already purchased car A. What is the oil consumption? Comparing to car B, which one is better?
		2. My tire is broken. Can this problem be fixed? How much it will cost me?
Topic 2: Experience sharing	Friend, Activities, Together, Life, Arrive, Place	1. My friend and his family came to my city during the weekend. I drive them around the city. We had a lot of fun together.
		2. It is snowing today! I am so lucky to have you by my side. Life is short but my love for you will never die.

In general, the discussion threads in my research context cover two topics: 1) information acquisition, driven by users' desire to search for information about products; 2) experience sharing, rising from users' tendency to share feelings related to product usage experience. The result shows that most threads (21,196) serve for the information acquisition function, whereas 7,818 threads are related to experience sharing, reflecting that the platforms are mainly used for knowledge or information exchange about automobiles. I code $Topic_j$ as a dummy variable to indicate the topic category of thread *j* and the variable is equal to 1 if the topic of the thread *j* is about information acquisition.

3.3.2. User Interaction

Dependency. Previous works have suggested that text message dependency, the overreliance on text-based online discussion, is driven by the need for interpersonal communication (Igarashi et al. 2008, Hayashi and Blessington 2020). In the online information diffusion setting, sequential dependency among online users exists, and the previous users' behavior affects the latter users' behavior (Dolatabadi et al. 2020). Therefore, I quantify user interaction by the dependency among user reply similarities over time. I apply *exponential smoothing*, a method that is widely used in Finance and Operational Research (Brown and Meyer 1961, Gardner 1985, Kalekar 2004), to calculate *content dependency* and *sentiment dependency*, respectively. Exponential smoothing is a technique that can capture the pattern of all historical data and is used for data forecasting (Brown and Meyer 1961, Tiao and Xu 1993). It is a model that indicates how well an individual learns (Brown and Meyer 1961).

I fit the content dependency and sentiment dependency using the series of similarity values between adjacent replies. I argue that the incoming user can learn from previous user interactions, with a focus on the similarity pattern between replies. To further illustrate, suppose that thread *j* has *n* replies, the simple smoothing functions of dependency can be modelled as follows:

$$Jaccar\widehat{d(A_{\jmath,1}}\,, A_{\jmath,0}) = Jacaard(A_{j,1}\,, A_{j,0}) \tag{3}$$

$$Jaccard\widehat{(A_{\jmath,\imath}}\,, A_{\jmath,\imath-1}) = \ \alpha Jacaard(A_{j,i}\,, A_{j,i-1}) + (1 - \ \alpha)Jaccard(\widehat{A_{\jmath,\imath-1}}\,, A_{\jmath,\imath-2}) \tag{4}$$

$$Cos\imath ne\widehat{(B_{\jmath,1}}\,, B_{\jmath,0}) = Cosine(B_{j,1}\,, B_{j,0}) \tag{5}$$

$$Cos\imath ne\widehat{(B_{\jmath,\imath}}\,, B_{\jmath,\imath-1}) = \ \beta Cosine(B_{j,i}\,, B_{j,i-1}) + (1 - \beta)Cos\imath ne(\widehat{B_{\jmath,\imath-1}}\,, B_{\jmath,\imath-2}) \tag{6}$$

Where *i=3,4,…,n* and $Jaccar\widehat{d(A_{\jmath,\imath}}\,, A_{\jmath,\imath-1})$ and $Cos\imath ne\widehat{(B_{\jmath,\imath}}\,, B_{\jmath,\imath-1})$ indicate the estimated content and sentiment similarity between *i*th reply and *(i-1)*th reply, respectively. The inverse smoothing factor, i.e., $1 - \alpha\ (1 - \beta)$, which is ranging from 0 to 1, is viewed as the thread-level content (sentiment) dependency on the previous information of the similarities between user replies. A greater value of dependency indicates that the similarity patterns of latter replies rely more on the similarity patterns of earlier replies (instead of recent patterns) in the thread. For each thread, I split the data into training, validation, and testing set, with the commonly used and well-performed ratio of 70:15: 15 (Akram et al. 2015, Lim et al. 2016). I apply the walk forward approach to have overfitting control and find the best estimate of the smoothing factor. Table 5 reports the average error over threads while calculating dependency, indicating good fitting and prediction performance.

Table 5. Average Error over Threads of Validation and Testing Set

	Content Coherence		Sentiment Coherence	
	Validation	Testing	Validation	Testing
RMSE	0.034	0.035	0.116	0.117
RMSLE	0.031	0.032	0.074	0.075
Note: RMSE means root mean square error; RMSLE means root mean squared logarithmic error.				

Regarding the descriptive statistics of dependency shown in Table 2, on average, the dependencies of both content and sentiment are high (0.86 and 0.76). The results indicate that users rely much on the similarities of previous replies when contributing new ones to the discussion. In other words, users are likely to observe and learn from previous user replies and follow those patterns. Interestingly, content dependency is on average 13.2% more than sentiment dependency. It suggests that users depend more on previous reply patterns when reading content relevant information whereas relying less on exiting user interactions when viewing emotional text. One explanation may be due to the possible substitution relationship between information coherence and dependency. When information coherence is high (the average of sentiment coherence is 0.92), users may easily perceive the consistently emotional pattern in the replies without digging into user interactions. While the relatively low content coherence (i.e., 0.11) motivates users to learn from the useful information in the previous replies. Another reason might be related to the discussing environment in my study. That is, users in thread discussion are mostly acquaintances with common interests. When participating in an online forum, the primary goal for users might be finding relevant content they are interested in. Thus, they will put more effort into the user interactions to find useful information, while less attention on emotion regulations.

Table 6. Tests on the Differences of Information Coherence and Dependency between Topics

	Information Acquisition	Experience Sharing	T-test[1]	P-value
	Mean (N1=21,196)	Mean (N2=7,818)		
$CONC_j$	0.115	0.107	-15.97	<0.001
$SENC_j$	0.939	0.863	-64.90	<0.001
$COND_j$	0.841	0.905	17.84	<0.001
$SEND_j$	0.732	0.822	20.90	<0.001
Note[1]: I also apply Wilcoxon signed-rank test without normal assumption. Results are consistent.				

I further compare the information coherence and the dependency of threads between the two underlying topics, i.e., information acquisition and experience sharing. Table 6 shows significant differences. Compared to the information coherence in the discussion about experience sharing, the discussion related to information acquisition has a higher level of content and sentiment coherence. It suggests that users are more concentrated on the content when discussing automobile features or problems. Meanwhile, their replies are relatively consistent in terms of sentiment (mostly neutral). In contrast, users rely more on the previous interactions among replies in the threads of experience sharing. The dependency of the threads about experience sharing is higher than that of information acquisition threads. This is probably because users care more about the patterns of previous replies in the threads where others share personal experiences. As a response, they add replies by referring to the similarities of the previous discussions and show emotional support. However, users tend to keep unique and express their own opinions professionally in a problem discussion, leading to lower dependency on the previous replies in the threads of information acquisition.

Original Poster's Replies. Another way of user interaction is the reciprocity indicated by the original poster's participation. I operationalize original poster's participation, $Original_j$, as log-transformed number of replies generated by the original poster in a thread discussion. From the perspective of the original poster, finding useful information or searching for emotional support are two reasons that motivate the generation of the original post. Thus, I consider the original poster has a strong tendency to interact with others. Other users who participate in the thread might be motivated by this reciprocity effect (Lakhani and Von Hippel 2004). As a group leader might facilitate further discussion (Herring and Nix 1997), I expect the leader of the thread discussion, namely, the original poster, her participation behavior might also influence information diffusion. On average, an original poster interacting with other participants approximately twice (i.e., 1.71) during the discussion (see Table 2).

I further compare the values of information coherence between whether the original poster has replied at least once in the thread discussion or not. The results in Table 7 suggest that, surprisingly, the information coherence is higher when the original poster does not participate in the thread discussion. Specifically, content coherence is approximately 0.9% higher with the absence of original poster's participation, whereas sentiment coherence increases by approximately 6.4% if the original poster does not reply. One possible explanation is that the original poster's reply may disrupt the existing coherence pattern or bring in new information sometimes.

Table 7. Tests on the Differences of Information Coherence between Original Poster's Reply

	Without Original Poster Reply	With Original Poster Reply	T-test[1]	P-value
	Mean (N1= 6,495)	Mean (N2= 22,519)		
$CONC_j$	0.114	0.113	2.024	<0.05
$SENC_j$	0.964	0.906	55.08	<0.001
Note[1]: I also apply Wilcoxon signed-rank test without normal assumption. Results are consistent.				

3.4. Duration and Breadth of Information Diffusion

The outcome of interest is information diffusion. I measure the dependent variables from two different perspectives – duration and breadth of information diffusion (Yang and Counts 2010, Zhang and Peng 2015). To be specific, I measure them as the count of active days when a thread has received at least one reply and the total number of replies that a thread has received, respectively. While empirical evidence has associated information coherence with reply volume (Weger and Aakhus 2003, Chen and Berger 2013), to the best of my knowledge, no research has linked information coherence to the duration of discussion.

3.5. Control Variables

I control for a couple of factors related to thread and user characteristics. Regarding thread-related factors, I include the log-transformed number of views as a proxy for the overall popularity of a thread. I also include a binary variable which indicates whether the thread is stated as valuable by the platform. I also control the length of the original post and the average length of all posts to capture user effort in the discussion. In addition, I consider the characteristics of the original poster in each thread. In particular, I control for the total number of threads the original poster in the focal thread (i) participated and (ii) started on the discussion platform during the observation window. I also control for two monthly sales ranks in the previous month before each thread starts: the rank among all types of automobiles and the SUV rank. Moreover, I include time dummies of the original post to capture unobserved events that might have happened every month. A dummy variable was incorporated to control the difference of user base from the two platforms. Besides, I included all the moderators in the main model as well as extension analyses. For the models using the number of replies as the outcome, I also control for the number of active days (i.e., duration).

4. Research Design and Results

4.1 Models and Identification Strategy

I evaluate the influence on information diffusion duration by survival analysis (i.e., Cox-proportional hazards model). It assumes that there is an unspecified baseline hazard $h_o(t)$, and variables that have a proportional effect on the underlying likelihood. Thus, I can capture how my variables of interests, i.e., content coherence and sentiment coherence, affect the likelihood that a thread will end given that the thread has lasted until a certain time point. Here, I define the end of a thread as when there is no reply beyond the 5 days since the last reply. The 5-days is determined because it is above the 99 percentile of the reply differences between any two replies in the data. Therefore, the right censoring happens if I can observe a reply within 5 days before the end of the observation window. The model is specified as follows:

$$h_j(t) = \ h_{oj}(t) \exp\left(\gamma_1 CONC_j + \gamma_2 CONC_j{}^2 + \gamma_3 SENC_j + \gamma_4 SENC_j{}^2 + \tau_j + \mu_j + \rho_j\right) \qquad (7)$$

where $h_{oj}(t)$ is the baseline hazard of thread $j.$ τ_j, μ_j, ρ_j represent time-, thread-, and user-level control effects in thread j respectively. The positive coefficients of γ indicate that the thread is more likely to change the status (end the thread) with an increase of information coherence measures, indicating a shorter discussion duration. I also consider the potential nonlinear effects and include the squared terms $CONC_j^2$ and $SENC_j^2$ in the Equation (7) to examine whether users' favor of consonance or feeling of redundancy drives the marginal effect of information coherence.

To estimate the impact on number of replies, I ran the negative binomial regression as $Replies_j$ in my data has significant overdispersion (i.e., likelihood-ratio test, p<0.01). Therefore, I model $Replies_j$ as a function of information coherence as follows:

$$E\left(Replies_j | \boldsymbol{x_j}, \epsilon_j\ \right) = \exp\left(\beta_1 CONC_j + \beta_2 CONC_j{}^2 + \beta_3 SENC_j + \beta_4 SENC_j{}^2 + \tau_j + \mu_j + \rho_j + \epsilon_j\right) \ (8)$$

where the number of replies follows the negative binomial distribution and exp (ϵ_j) captures the unobserved heterogeneity.

One concern here is that information coherence is endogenously formed, such that the information diffusion process might also affect information coherence along the thread. For example, as more users participate in an online discussion, they might be more willing to show their uniqueness through posting, which might lower the information coherence of the thread. To address the potential endogeneity problem, I use lagged content coherence and sentiment coherence from other threads (Bartik 1991) and construct my instrument variables as follows:

$$CONCIV_j = Avg\ (CONC_{i,-j}) \tag{9}$$

$$SENCIV_j = Avg\ (SENC_{i,-j}) \tag{10}$$

where $CONCIV_j$ and $SENCIV_j$ measure the simple average of content coherence and sentiment coherence in the other threads that the users in the focal thread j have joined in the recent past. Here $CONC_{i,-j}$ and $SENC_{i,-j}$ are the content and sentiment coherence of the most recent (other) thread that user *i* has joined in the past. I only use the data from the users who have previous participation in the thread discussion. The instrument variables are believed as valid because a user may share a similar preference on the pattern of information coherence when she decides to participate in the discussion. Previous research has indicated that viewing the information in other threads might provide insights on the information that the user processes in the focal thread (Kane and Ransbotham 2016). Therefore, the information coherence from the other thread that the user has recently participated in might be associated with the coherence pattern in the focal thread. However, an individual's preference for thread participation is less likely to affect the overall diffusion of the focal thread discussion. I employ a Poisson regression with endogenous regressors (i.e., Poisson GMM) to estimate the relationships. In my main analysis, I explain the effects of information coherence using the results from the Poisson GMM, although I also show the estimation results from the standard negative binomial model. I only report the results in the extension analyses using the Cox model and Poisson GMM.

4.2 Main Effect of Information Coherence

Table 8 shows the main results. In terms of the linear relationships, the results show that when holding the other variables constant, increasing $CONC_j$ by one unit increases the likelihood of ending a thread discussion by 6.9% (exp (0.067) =1.069). In other words, higher content coherence will lead to a shorter discussion. On the other side, a unit increase of $CONC_j$ increases the number of replies by 2.7% (exp (0.027) = 1.027). The results suggest a problem-solving phenomenon. That is, users like to solve problems efficiently in online discussion (i.e., threads have shorter duration but more replies) if the content is more relevant across the discussion. However, increasing $SENC_j$ by one unit decreases the likelihood of ending a thread discussion by 6.7% (exp (-0.069) = 0.933). The threads will be active for longer time with higher sentiment coherence. On the other side, the number of replies decreases by 7.1% (exp (-0.074) = 0.929) with the increase of $SENC_j$. Under this circumstance, users' tendency is not to solve problems efficiently (i.e., threads have longer duration but fewer replies) if the sentiment is more consistent across the discussion. The results suggest that users take different viewpoints about the coherence of content and sentiment in online discussion. Higher content coherence from previous replies reflects an informative and relevant discussion context, which speeds up the problem-solving process by attracting more reply volume within a shorter duration. However, higher sentiment coherence indicates consistently positive, neutral, or negative opinions in the previous discussion, which might enhance users' feeling of information redundancy. As a result, users might lose interest to intensively participate in the discussion, while some occasional posts may be added during the thread diffusion.

Table 8. Results of Duration and Reply Volume

	Cox-proportional Hazard		Poisson GMM		Negative Binomial	
	DV= Duration		DV= Replies		DV= Replies	
	Linear	Square	Linear	Square	Linear	Square
$CONC_j$	0.067***	0.091***	0.027**	0.070***	0.045***	0.060***
	(5.67e-03)	(7.74e-03)	(1.11e-02)	(1.81e-02)	(2.93e-03)	(3.62e-03)
$CONC_j^2$		-0.007***		-0.004*		-0.003***
		(1.58e-03)		(2.17e-03)		(4.80e-04)
$SENC_j$	-0.069***	0.004	-0.074***	-0.346***	-0.004	-0.091***
	(8.33e-03)	(1.30e-02)	(2.05e-02)	(5.50e-02)	(4.04e-03)	(6.31e-03)
$SENC_j^2$		0.036***		-0.156***		-0.044***
		(4.70e-03)		(2.68e-02)		(2.45e-03)
$Duration_j$			0.334***	0.319***	0.362***	0.359***
			(5.56e-03)	(6.45e-03)	(3.71e-03)	(3.70e-03)
Constant			-0.598***	-0.415***	-1.278***	-1.187***
			(1.25e-01)	(1.35e-01)	(5.49e-02)	(5.49e-02)
Time Controls	Yes	Yes	Yes	Yes	Yes	Yes
Thread Controls	Yes	Yes	Yes	Yes	Yes	Yes
Users Controls	Yes	Yes	Yes	Yes	Yes	Yes
Endogeneity Correction			Yes	Yes		
No. of Obs.	29,014	29,014	29,014	29,014	29,014	29,014
Log likelihood	-262390	-262349			-110679	-110508
Pseudo- R^2					0.211	0.212

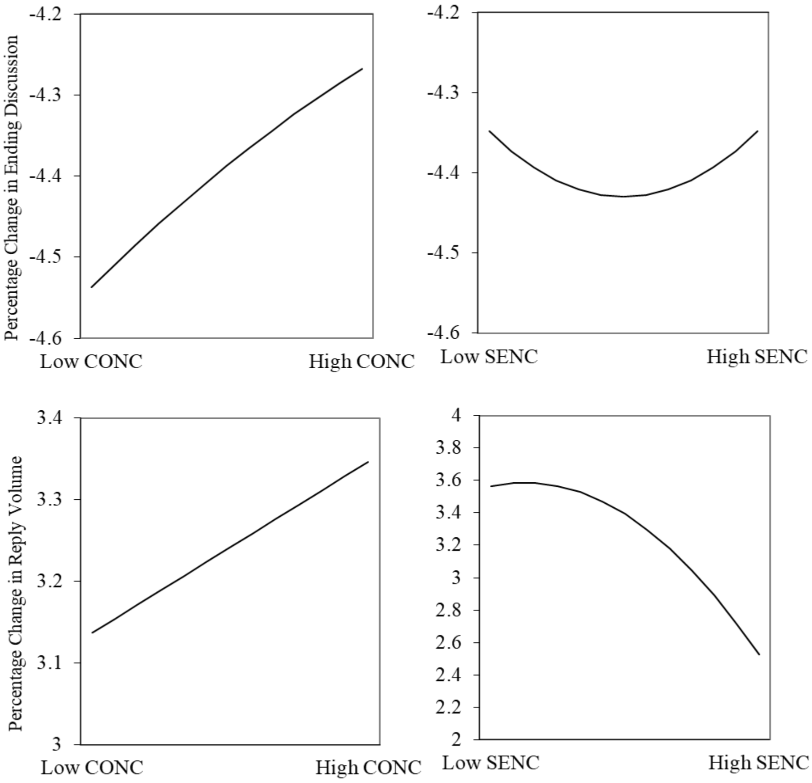
**** p<.01, ** p<.05, * p<.1*

I further explore the non-linear relationships of information coherence and diffusion measures. All the square term coefficients are significant except for the coefficient of CONC^2 on generating replies.

To understand the influence clearly, I plot the curvilinear effects in Figure 2. The positive effect of $CONC_j$ on ending thread discussion is subject to diminishing returns (-0.007, p<0.01), and the effect of $CONC_j$ is mostly positive in my data range as it turns negative when $CONC_j$ exceeds 99.8 percentile (-0.0908377/(2*-0.0067329) = 6.746)). This observation again confirms the strong positive effect of content relevance on reducing discussion duration. I don't find a curvilinear effect of content coherence on the number of replies (p-value > 0.05). Taken together, the effect of content coherence suggests that the threads with higher content coherence will last shorter but with more replies. The marginal effect of content coherence on ending the discussion indicates that the cognitive dissonance drives information diffusion under high content coherence, as consonance leads to an increase in discussion duration with the additional increase in content coherence.

The curvilinear relationship between $SENC_j$ and information diffusion suggests a nuanced pattern. When $SENC_j$ is at a low level (i.e., below 15 percentile), a higher $SENC_j$ leads to the lower probability of ending discussion (longer duration), and meanwhile, it boosts reply volume with a positive influence. In this case, users tend to actively engage in the discussion and contribute to longer duration and more replies. When $SENC_j$ is at a moderate level (i.e., around 15 to 40 percentile), an increase in $SENC_j$ increase discussion duration while decrease reply volume, which hinders the efficiency of solving the problem. When the sentiment coherence is extremely high, the threads last shorter with fewer replies. Taken together, the effect of sentiment coherence suggests that threads with moderately sentiment coherence have more replies in a longer duration. The negative marginal effect of sentiment coherence indicates that the cognitive load drives information diffusion under high sentiment coherence, as the feeling of redundancy leads to decreases in discussion duration and reply volume with the additional increase in sentiment coherence.

Figure 2. Relationship Between Information Coherence and Information Diffusion



5. Model Extensions and Robustness Checks

I extend main analysis by further exploring the moderating effects of topics and user interactions on information coherence. I also conduct some robustness checks to show that my findings are consistent across different model specifications and measurements.

5.1. The Moderating Effect of Topics

Users in the online discussion do not treat topics equally. They tend to participate in the discussion topic they are interested in. For example, Chaiken (1980) indicates that the importance of opinion judgement will influence the extent of response involvement. The topic from a non-consequential message has been shown to lead to effort minimization. Blau and Barak (2012) suggest that comparing to information diffusion in the non-sensitive topic, information diffusion in sensitive topic generate more participation and higher contribution quality. I speculate that discussion topic may moderate the relationship between information coherence and information diffusion.

As mentioned in the above section, I infer the topic of each thread based on its original post. Based on LDA model results, the topic of each thread in my data belongs to either information acquisition or experience sharing. The variable $Topic_j$ is equal to 1 if the topic of thread j is about information acquisition. Table 9 reports the results. The positive and significant coefficients of $CONC_j \times Topic_j$ and $SENC_j \times Topic_j$ in the survival model suggest that high content and sentiment coherence will end the thread discussion faster when the thread is about information acquisition (vs. experience sharing). On the other hand, the effect of $CONC_j$ on replies decreases if the topic is about information acquisition. Therefore, users tend to close the topic quickly with fewer replies in the discussion about information acquisition when the previous consensus has already been achieved. In contrast, users will engage more actively in the thread about experience sharing with a longer duration and more replies if the discussion is coherent. It suggests that users are more sensitive to redundancy when the discussion is about information acquisition.

Table 9. The Moderating Effect of Topics

Model	(1)	(2)
	Cox-proportional Hazard	Poisson GMM
	DV= Durations	DV = Replies
$CONC_j$	0.031**	0.095***

	(1.31e-02)	(2.05e-02)
$SENC_j$	-0.135***	-0.051**
	(1.30e-02)	(2.32e-02)
$CONC_j \times Topic_j$	0.045***	-0.090***
	(1.44e-02)	(2.49e-02)
$SENC_j \times Topic_j$	0.098***	-0.029
	(1.49e-02)	(1.79e-02)
$Topic_j$	0.164***	-0.129***
	(1.74e-02)	(1.86e-02)
$Duration_j$		0.332***
		(5.65e-03)
Constant		-0.603***
		(1.28e-01)
Time Controls	Yes	Yes
Thread Controls	Yes	Yes
Users Controls	Yes	Yes
Endogeneity Correction		Yes
No. of Obs.	29,014	29,014
Log likelihood	-262363	

**** p<.01, ** p<.05, * p<.1*

5.2. The Moderating Effect of User Interactions

5.2.1. Does the Effect of Information Coherence Differ by Dependency?

Dependency measures one type of user interaction. Wang et al. (2010) show that when dependency exists among inter-organizational partners, the damage caused by incoherence on exchange performance is mitigated. A similar result has been found in the context of online review. Dolatabadi et al. (2020) suggest that previous decisions in online reviews can influence the current decision of the focal user. They further indicate that the review popularity, the polarity is extracted based on review messages by the NLP tool, is less likely to be positive if it follows positive reviews. Therefore, I speculate that dependency, that is, the extent to which the later post similarity relies on previous ones, might moderate the effect of information coherence on information diffusion.

I operationalize $COND_j$ and $SEND_j$ of thread j by calculating the smoothing parameters in exponential smoothing based on similarity patterns in $CONC_j$ and $SENC_j$ respectively, which indicate how well users learn from previous similarity patterns in the thread discussion (Brown and Meyer 1961). Higher dependence suggests that users rely more on the information in previous user interactions. The estimation results of the moderating effects are presented in Model (1) and Model (2) in Table 10.

The coefficients of $COND_j$ and $SEND_j$ suggest that when dependency increase, the probability of ending the thread discussion decrease (-0.063, p<0.01) whereas the number of replies increases (0.087, p<0.01; 0.024, p<0.01). These observations indicate that higher dependency motivates user engagement directly (i.e., more replies in a longer duration). Regarding the interaction terms, the coefficient of $CONC_j \times COND_j$ is positive and significant (i.e., 0.026, p<0.01) whereas the coefficient of $SENC_j \times SEND_j$ is negative and significant (i.e., -0.031, p<0.01), suggesting that under a thread with high dependency, the effects of $CONC_j$ and $SENC_j$ on thread duration become even stronger (the effect is more positive from content coherence, while the effect of sentiment coherence is more negative). The moderating effect of $COND_j$ negatively influences the impact of $CONC_j$ on the number of replies (i.e., -0.029, p<0.01), suggesting an increasing feeling of redundancy under high dependency, which reduces the marginal effect of $CONC_j$ on generating more replies.

Taken together, I find that the moderating effects of $COND_j$ on $CONC_j$ demotivates user replies and discussion duration. The reason might be that with higher dependency (vs. lower dependency), users learn the information in content faster, which helps to find a wanted answer in a shorter period as well as more sensitive to information redundancy. The redundancy might result from the substitution information in content coherence and dependency when the dependency is high.

Table 10. The Moderating Effect of User Interaction

Model	(1)	(2)	(3)	(4)
	Cox-proportional Hazards	Poisson GMM	Cox-proportional Hazards	Poisson GMM
	DV = Duration	DV= Replies	DV = Duration	DV= Replies
$CONC_j$	0.074*** (6.00e-03)	0.024 (1.50e-02)	0.068*** (7.64e-03)	-0.023 (3.23e-02)
$SENC_j$	-0.062*** (8.38e-03)	-0.071*** (2.23e-02)	0.044*** (1.40e-02)	-0.272*** (3.93e-02)
$CONC_j \times COND_j$	0.026*** (5.15e-03)	-0.029*** (1.09e-02)		
$SENC_j \times SEND_j$	-0.031*** (7.04e-03)	-0.002 (6.75e-03)		
$CONC_j \times Original_j$			-0.003 (4.06e-03)	0.017* (9.87e-03)
$SENC_j \times Original_j$			-0.060*** (5.71e-03)	0.068*** (9.37e-03)
$COND_j$	-0.063*** (6.25e-03)	0.087*** (5.71e-03)		
$SEND_j$	-0.012* (6.55e-03)	0.024*** (4.33e-03)		

$Original_j$			-0.116**	-0.766***
			(1.74e-02)	(7.79e-02)
$Duration_j$		0.327***		0.325***
		(5.58e-03)		(6.05e-03)
Constant		-0.532***		-0.785***
		(1.26e-01)		(1.32e-01)
Time Controls	Yes	Yes	Yes	Yes
Thread Controls	Yes	Yes	Yes	Yes
Users Controls	Yes	Yes	Yes	Yes
Endogeneity Correction		Yes		Yes
No. of Obs.	29,014	29,014	29,014	29,014
Log likelihood	-262313		-262333	

*** *p<.01*, ** *p<.05*, * *p<.1*

5.2.2. Does the Effect of Information Coherence Differ by Original Poster's Replies?

Another measurement of user interaction is the original poster's replies. I expect that replies from the original poster play an important role in moderating the relationship between information coherence and information diffusion. By interviewing managers of an accounting firm, Cross and Sproull (2004) indicate that managers are more likely to receive help from those they have provided valuable help to. In the online community, reciprocity itself is a reason why users participate in discussions online (Lakhani and Von Hippel 2004), even when the participants are newcomers (Joyce and Kraut 2006). It has been shown that when taking reciprocity into the modelling consideration, the model can better fit the discussion pattern therein (Faraj and Johnson 2011, Johnson et al. 2014). Thus, I speculate that users are more likely to participate in the online discussion if the original poster replies. In addition, the original posters' replies moderate the relationships between information coherence and information diffusion. In the analyses, I code $Original_j$ as log-transformed number of replies generated by the original poster.

The results of Model (3) and Model (4) in Table 10 show that when the original poster's reply increases, the likelihood of ending the discussion decreases (-0.116, p<0.05). However, the probability of generating more replies also decreases (-0.766, p<0.01). It indicates that the threads with the original poster's participation have fewer replies and a shorter duration. The reason could be that when there are few posts in the thread, the original poster actively comments to motivate more participation. It is also likely that the problem is easily and quickly addressed by other users, thus the original poster comment in the thread for appreciation.

The coefficient of $SENC_j \times Original_j$ is negative on the probability of ending the discussion (i.e., -0.060, p<0.01). Based on the derivative of Model (3) in Table 10 with respect to $SENC_j$, I can observe that the effect of $SENC_j$ on ending the discussion first increases with more replies from the original poster. When $Original_j$ exceeds the 35 percentile (i.e., 0.044/0.06 = 0.733), this positive effect of $SENC_j$ turns into negative. While the coefficient of $SENC_j \times Original_j$ on number of replies is positive, and the derivative of Model (4) in Table 10 with respect to $SENC_j$ suggests that the effect of sentiment coherence on replies is mostly negative with the increase of replies from the original poster in the data range as it turns positive after $Original_j$ exceeds the 92 percentile (i.e., 0.272/0.068 = 4). Therefore, the number of replies in the threads decreases with the sentiment coherence upon the increase of original poster replies. The coefficients of $CONC_j \times Original_j$ in both Model (3) and Model (4) in Table 10 are not statistically significant (i.e., p>0.05), suggesting the reciprocity from original posters only affects the relationship between sentiment coherence and information diffusion. Besides, moderating effect of the original poster's replies on sentiment coherence positively influences discussion duration and replies, suggesting that users are willing to diffuse information when consensus in sentiment is reached with the interaction from the original poster.

5.3. Robustness checks

I conduct a series of robustness checks to test the main findings. 1) I show that the effects of content and sentiment coherence still hold with alternative model specifications. I re-evaluate duration using a negative binomial model and replies using an endogenous ordinary least squares model estimated by two-stage least squares (2SLS). I also log-transformed the number of replies in the 2SLS model. The model results are consistent with the main analysis and are shown in Table 11. 2) I consider different measurements regarding information diffusion. I use the day difference between the first and the last posts to be an alternative measure of duration and evaluate the likelihood of ending the discussion using semi-parametric proportional modelling (i.e., Cox model). I also directly evaluate the problem-solving efficiency, measured by replies divided by active comment days. The ordinary least squares model is estimated by GMM to address potential endogeneity issue. The results echo that content coherence increases problem-solving efficiency, whereas sentiment coherence decreases problem-solving efficiency (see Table 12).

Table 11. Alternative Model Specifications

	Negative Binomial		2SLS	
	DV= Duration		DV= Replies	
	Linear	Square	Linear	Square
$CONC_j$	-0.092***	-0.117***	0.056***	0.168***
	(5.16e-03)	(6.38e-03)	(1.10e-02)	(1.95e-02)
$CONC_j^2$		0.009***		-0.022***
		(1.22e-03)		(3.82-03)
$SENC_j$	0.090***	-0.016	-0.128***	-0.431***
	(6.96e-03)	(1.11e-02)	(1.65e-02)	(4.74e-02)
$SENC_j^2$		-0.057***		-0.199***
		(4.66e-03)		(2.67e-02)
Constant	-2.88***	-2.80***	-1.11***	-0.516***
	(9.38e-02)	(9.42e-02)	(4.75e-02)	(7.91e-02)
Time Controls	Yes	Yes	Yes	Yes
Thread Controls	Yes	Yes	Yes	Yes
Users Controls	Yes	Yes	Yes	Yes
Endogeneity Correction			Yes	Yes
No. of Obs.	29,014	29,014	29,014	29,014
Log likelihood	-67969	-262349		
R^2			0.86	0.84

**** p<.01, ** p<.05, * p<.1*

Table 12. Alternative Measurements of Information Diffusion

	Cox-proportional Hazards		GMM	
	DV= Day Difference		DV= Efficiency	
	Linear	Square	Linear	Square
$CONC_j$	0.062***	0.078***	3.51***	6.14***
	(5.82e-03)	(7.66e-03)	(3.47e-01)	(1.26e+00)
$CONC_j^2$		-0.005***		-1.02**
		(1.58e-03)		(4.14e-01)
$SENC_j$	-0.070***	-0.033***	-6.41***	2.87*
	(8.12e-03)	(1.28e-02)	(4.94e-01)	(1.46e+00)
$SENC_j^2$		0.019***		6.06***
		(4.86e-03)		(9.01e-01)
Constant			-19.28***	-28.48***
			(1.30e+00)	(3.99e+00)
Time Controls	Yes	Yes	Yes	Yes
Thread Controls	Yes	Yes	Yes	Yes
Users Controls	Yes	Yes	Yes	Yes
Endogeneity Correction			Yes	Yes
No. of Obs.	29,014	29,014	29,014	29,014
Log likelihood	-261436	-261422		
R^2			0.34	0.01

**** p<.01, ** p<.05, * p<.1*

6. Discussion

In this research, I set out to examine whether and how the effect of information coherence influences information diffusion. Using the data from popular online automobile discussion platforms in China, I empirically fit the patterns of coherence along the sequence of user posts and estimate the dependency by a data forecasting method. I measure the textual information by two dimensions, content and sentiment, to capture the content relevance and the sentiment consistency among the user posts. I have the following key findings. First, the heterogeneous effect of information coherence does exist. Higher content coherence leads to more replies in a shorter duration; however, higher sentiment coherence leads to fewer replies in a longer duration, which shows the opposite outcomes. The results from nonlinear relationships suggest that while the effect of content coherence on duration is mostly positive in the data range, a curvilinear relationship between sentiment coherence and information diffusion exists, such that a moderate level of sentiment coherence leads to more replies and longer duration. In terms of theories, I provide suggestive evidence that cognitive consonance drives the marginal effect of content coherence on information diffusion, whereas redundancy drives the marginal effect of sentiment coherence on information diffusion. I also found that the moderating effect of topic on information acquisition and dependency on content coherence drives by redundancy, which results in fewer replies and shorter duration in a thread discussion. The moderating effect of the original poster's replies on sentiment coherence drives by cognitive consonance, which results in more replies and longer duration in a thread discussion. These moderating results further shed light on the heterogeneous effect of information coherence.

6.1 Theoretical and Managerial Implications

This research contributes to the literature in several ways. First, I extend the literature on the antecedent of online discussion. Since the emergence of microblogging, incoherence is perceived as a dominant threat to understanding online discussion (Honey and Herring 2009, Abbasi et al. 2018), and researchers have concentrated particularly on the social network in online reviews and social media (Huang et al. 2016, Stella et al. 2018, Burbach et al. 2019). My work is built on research that re-evaluates the magnitude of coherence in online discussions. For example, Stromer-Galley and Martinson (2009) conduct a comparison study in online chat rooms to investigate coherence when people synchronously discuss online. They find evidence of a relatively high coherence among political discussion, which contradicts previous findings that synchronous online discussion suffers from low coherence (Weger and Aakhus 2003). In the context of online discussion where users are self-motivated to participate, I causally identify and empirically show that information coherence is also an important antecedent of online discussion. The results suggest that the heterogeneous and nonlinear effect of information coherence does exist, which might count for the previous contradictory findings towards the effect of coherence. In addition, I also show how this effect of information coherence changes with the moderating of the topic, user interactions, and original poster replies, which further sheds light on the heterogeneity of information coherence.

Second, my research adds to the literature on user interaction by providing an innovative perspective of a data forecasting method that describes dependency among user interactions in a thread. The dependency is built on the smoothing factor from exponential smoothing. It offers a sophisticated way to quantify the extent to which, on average, the similarities of the latter posts rely on the similarity patterns of earlier posts in the thread. Given the endogenous history and exogenous forecasts, I implicatively limit the potential endogenous issue between information coherence and dependency. Unlike the previous research that focuses on pairwise user interaction in social networks (Viswanath et al. 2009, Wang et al. 2018), the dependency measurement in this research captures the interdependent pattern among thread-level user interaction.

Third, I shed new light on the tension between cognitive consonance and redundancy in online discussion. I first explore the marginal effect of information coherence measured by the square term of content coherence and sentiment coherence. I find that cognitive consonance drives the marginal effect of content coherence on information diffusion, whereas redundancy drives the marginal effect of sentiment coherence on information diffusion. Leveraging the moderation analyses, I further show that moderating effect of the topic of information acquisition and dependency on content coherence drives by redundancy, whereas moderating effect of the original poster's replies on sentiment coherence drives by cognitive consonance.

The results have important managerial implications to facilitate platforms' online discussion as well as companies' product management. Content relevance is shown to be more important than sentiment consistency to increase replies in a shorter period. However, while a content-relevant discussion may help resolve users' problems quickly, there is a cost to encourage further discussion. Instead, sharing experience might create fun and thus encourage interaction and engage users better. This study brings insights into the strategy designs for online discussion platforms and business managers to improve user engagement (i.e., longer duration and more replies). Due to the heterogeneity in information coherence, the effect of content and sentiment coherence under the same rules might works differently on information diffusion. Therefore, administrators should be carefully designing the rules of online discussion. The moderating effect of user interaction suggests that user interaction cannot always lead to more engagement. Encouraging higher dependency when the content coherence level is low is recommended to the platform administrators. Besides, our results indicate that it is important for product companies to detect threads with original poster replies frequently, especially for threads with negative sentiment before these threads go viral. An active original poster could make the complaint thread remain active for long period with increasing replies to complaints.

6.2 Limitations and Future Research

Despite the robustness of my empirical evidence, this research has limitations that suggest future work. First, empirical study in this research takes advantage of the fact that most users in online forums read posts based on the sequence of its display (which is in time order). If more dataset is available, I could explore how my findings can be generalized to the context of online reviews and social media. Second, the measure of information coherence is constructed based on the textual information, even though some posts also contain image information. Besides, there is an emerging trend of online platforms embedding other features such as short-formed video. Although limited by the data, it would be interesting to discover how those features could shift my findings. Third, because of the limitation in the dataset, I could not observe the impact of information coherence on sales of corresponding cars. Future research should extend my exploratory study with a particular focus on the marketing impact of online information coherence.

7. References

Abbasi A, Zhou Y, Deng S, Zhang P (2018) Text Analytics to Support Sense-Making in Social Media: A Language-Action Perspective. MIS Quarterly 42(2): 427-464.

AbuSafiya M (2020) Measuring Documents Similarity using Finite State Automata. 2020 2nd International Conference on Mathematics and Information Technology (ICMIT) (IEEE), 208-211.

Akram S, Javed MY, Qamar U, Khanum A, Hassan A (2015) Artificial neural network based classification of lungs nodule using hybrid features from computerized tomographic images. Applied Mathematics & Information Sciences 9(1):183.

Alexander Jr CN (1964) Consensus and mutual attraction in natural cliques a study of adolescent drinkers. American Journal of Sociology 69(4):395-403.

Ananthakrishnan UM, Li B, Smith MD (2020) A Tangled Web: Should Online Review Portals Display Fraudulent Reviews? Information Systems Research 31(3):950-971.

Aral S, Walker D (2011) Creating social contagion through viral product design: A randomized trial of peer influence in networks. Management science 57(9):1623-1639.

Aral S, Brynjolfsson E, Van Alstyne M (2007) Productivity effects of information diffusion in e-mail networks. ICIS 2007 Proceedings:17.

Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proceedings of the National Academy of Sciences 106(51):21544-21549.

Avramova Z, Wittevrongel S, Bruneel H, De Vleeschauwer D (2009) Analysis and modeling of video popularity evolution in various online video content systems: Power-law versus exponential decay. 2009 First International Conference on Evolving Internet (IEEE), 95-100.

Bartik TJ (1991) Who benefits from state and local economic development policies?

Barzilay R, Elhadad M (1999) Using lexical chains for text summarization. Advances in automatic text summarization:111-121.

Bhatia S, Mitra P (2010) Adopting inference networks for online thread retrieval. Proceedings of the AAAI Conference on Artificial Intelligence.

Bhattacharjee S, Das A, Bhattacharya U, Parui SK, Roy S (2015) Sentiment analysis using cosine similarity measure. 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS) (IEEE), 27-32.

Blau I, Barak A (2012) How do personality, synchronous media, and discussion topic affect participation? Journal of Educational Technology & Society 15(2):12-24.

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research 3(Jan):993-1022.

Bou-Franch P, Lorenzo-Dus N, Blitvich PG-C (2012) Social interaction in YouTube text-based polylogues: A study of coherence. Journal of Computer-Mediated Communication 17(4):501-521.

Brown RG, Meyer RF (1961) The fundamental theorem of exponential smoothing. Operations Research 9(5):673-685.

Burbach L, Halbach P, Ziefle M, Calero Valdez A (2019) Who Shares Fake News in Online Social Networks? Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, 234-242.

Chaiken S (1980) Heuristic versus systematic information processing and the use of source versus message cues in persuasion. Journal of personality and social psychology 39(5):752.

Chan J, Hayes C, Daly E (2010) Decomposing discussion forums and boards using user roles. Proceedings of the International AAAI Conference on Web and Social Media.

Chen Z, Berger J (2013) When, why, and how controversy causes conversation. Journal of Consumer Research 40(3):580-593.

Cialdini RB, Goldstein NJ (2004) Social influence: Compliance and conformity. Annu. Rev. Psychol. 55:591-621.

Cross R, Sproull L (2004) More than an answer: Information relationships for actionable knowledge. Organization Science 15(4):446-462.

De Beaugrande R-A, Dressler WU (1981) Introduction to text linguistics (Longman London).

De Choudhury M, Sundaram H, John A, Seligmann DD, Kelliher A (2010) " Birds of a Feather": Does User Homophily Impact Information Diffusion in Social Media? arXiv preprint arXiv:1006.1702.

Della Briotta Parolo P, Pan RK, Ghosh R, Huberman BA, Kaski K, Fortunato S (2015) Attention decay in science. arXiv e-prints:arXiv: 1503.01881.

Deng F, Siersdorfer S, Zerr S (2012) Efficient jaccard-based diversity analysis of large document collections. Proceedings of the 21st ACM international conference on Information and knowledge management, 1402-1411.

Deng S, Zhang P, Zhou Y (2011) Turning Unstructured and Incoherent Group Discussion into DATree: A TBL Coherence Analysis Approach. Proceedings of the International Conference on Information Systems, ICIS 2011, Shanghai, China, December 4-7, 2011.

Dewan S, Ho Y-J, Ramaprasad J (2017) Popularity or proximity: Characterizing the nature of social influence in an online music community. Information Systems Research 28(1):117-136.

Dolatabadi M, Fadardi JS, Kahani M, Karshki H (2020) Cognitive sequential dependencies in the wild: Sentiment analysis approach.

Ehrlich D, Guttman I, Schönbach P, Mills J (1957) Postdecision exposure to relevant information. The journal of abnormal and social psychology 54(1):98.

Faraj S, Johnson SL (2011) Network exchange patterns in online communities. Organization science 22(6):1464-1480.

Festinger L (1957) A theory of cognitive dissonance (Stanford university press).

Freedman JL (1965) Confidence, utility, and selective exposure: A partial replication. Journal of personality and social psychology 2(5):778.

Fu T, Abbasi A, Chen H (2008) A hybrid approach to web forum interactional coherence analysis. Journal of the American Society for Information Science and Technology 59(8):1195-1209.

Gao B, Pavel L (2017) On the properties of the softmax function with application in game theory and reinforcement learning. arXiv preprint arXiv:1704.00805.

Gardner Jr ES (1985) Exponential smoothing: The state of the art. Journal of forecasting 4(1):1-28.

Givón T (1995) Functionalism and grammar (John Benjamins Publishing).

Godes D, Silva JC (2012) Sequential and temporal dynamics of online opinion. Marketing Science 31(3):448-473.

Guille A, Hacid H (2012) A predictive model for the temporal dynamics of information diffusion in online social networks. Proceedings of the 21st international conference on World Wide Web, 1145-1152.

Hagel J (1999) Net gain: Expanding markets through virtual communities. Journal of interactive marketing 13(1):55-65.

Hayashi Y, Blessington GP (2020) Excessive valuation of social interaction in text-message dependency: A behavioral economic demand analysis. The Psychological Record:1-9.

Hecking T, Chounta I-A, Hoppe HU (2016) Investigating social and semantic user roles in MOOC discussion forums. Proceedings of the sixth international conference on learning analytics & knowledge, 198-207.

Herring S (1999) Interactional Coherence in Cmc. Journal of Computer-Mediated Communication 4(4).

Herring SC (1999) Interactional coherence in CMC. Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers (IEEE), 13 pp.

Herring SC, Nix C (1997) Is “serious chat” an oxymoron? Academic vs. social uses of Internet Relay Chat. American Association of Applied Linguistics, Orlando, FL, March 11.

Hill K, Fitzgerald R (2020) Student perspectives of the impact of COVID-19 on learning. All Ireland Journal of Higher Education 12(2):1-9.

Hoffman B (2010) “I think I can, but I'm afraid to try”: The role of self-efficacy beliefs and mathematics anxiety in mathematics problem-solving efficiency. Learning and individual differences 20(3):276-283.

Hoffman B, Schraw G (2009) The influence of self-efficacy and working memory capacity on problem-solving efficiency. Learning and Individual Differences 19(1):91-100.

Honey C, Herring SC (2009) Beyond microblogging: Conversation and collaboration via Twitter. 2009 42nd Hawaii International Conference on System Sciences (Ieee), 1-10.

Huang N, Hong Y, Burtch G (2016) Social network integration and user content generation: Evidence from natural experiments. MIS Quarterly (Forthcoming):17-001.

Ibrahim NF, Wang X, Bourne H (2017) Exploring the effect of user engagement in online brand communities: Evidence from Twitter. Computers in Human Behavior 72:321-338.

Igarashi T, Motoyoshi T, Takai J, Yoshida T (2008) No mobile, no life: Self-perception and text-message dependency among Japanese high school students. Computers in Human Behavior 24(5):2311-2324.

Im Y-H, Kim E-m, Kim K, Kim Y (2011) The emerging mediascape, same old theories? A case study of online news diffusion in Korea. New Media & Society 13(4):605-625.

Iribarren JL, Moro E (2009) Impact of human activity patterns on the dynamics of information diffusion. Physical review letters 103(3):038702.

Jaakonmäki R, Müller O, Vom Brocke J (2017) The impact of content, context, and creator on user engagement in social media marketing. Proceedings of the 50th Hawaii international conference on system sciences.

Jabr W, Mookerjee R, Tan Y, Mookerjee VS (2014) Leveraging philanthropic behavior for customer support: The case of user support forums. MIS quarterly 38(1):187-208.

Johnson SL, Faraj S, Kudaravalli S (2014) Emergence of Power Laws in Online Communities. Mis Quarterly 38(3):795-A13.

Joyce E, Kraut RE (2006) Predicting continued participation in newsgroups. Journal of Computer-Mediated Communication 11(3):723-747.

Kalekar PS (2004) Time series forecasting using holt-winters exponential smoothing. Kanwal Rekhi School of Information Technology 4329008(13):1-13.

Kane GC, Ransbotham S (2016) Research note—content and collaboration: an affiliation network approach to information quality in online peer production communities. Information Systems Research 27(2):424-439.

Kane GC, Johnson J, Majchrzak A (2014) Emergent life cycle: The tension between knowledge change and knowledge retention in open online coproduction communities. Management Science 60(12):3026-3048.

Kornfield R, Toma CL (2020) When do Online Audiences Amplify Benefits of Self-Disclosure? The Role of Shared Experience and Anticipated Interactivity. Journal of Broadcasting & Electronic Media:1-21.

Korolija N (2000) Coherence-inducing strategies in conversations amongst the aged. Journal of pragmatics 32(4):425-462.

Kuang L, Huang N, Hong Y, Yan Z (2019) Spillover effects of financial incentives on non-incentivized user engagement: Evidence from an online knowledge exchange platform. Journal of Management Information Systems 36(1):289-320.

La Fond T, Neville J (2010) Randomization tests for distinguishing social influence and homophily effects. Proceedings of the 19th international conference on World wide web, 601-610.

Lakhani KR, Von Hippel E (2004) How open source software works:“free” user-to-user assistance. Produktentwicklung mit virtuellen Communities (Springer), 303-339.

Lehmann J, Lalmas M, Yom-Tov E, Dupret G (2012) Models of user engagement. International conference on user modeling, adaptation, and personalization (Springer), 164-175.

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

Lim WT, Wang L, Wang Y, Chang Q (2016) Housing price prediction using neural networks. 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) (IEEE), 518-522.

Liu Y, Jiang C, Zhao H (2018) Using contextual features and multi-view ensemble learning in product defect identification from online discussion forums. Decision Support Systems 105:1-12.

Ma L, Krishnan R, Montgomery AL (2015) Latent homophily or social influence? An empirical analysis of purchase within a social network. Management Science 61(2):454-473.

Mason L, Ariasi N, Boldrin A (2011) Epistemic beliefs in action: Spontaneous reflections about knowledge and knowing during online information searching and their influence on learning. Learning and Instruction 21(1):137-151.

McNamara DS, Kintsch W (1996) Learning from texts: Effects of prior knowledge and text coherence. Discourse processes 22(3):247-288.

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Annual review of sociology 27(1):415-444.

Medin DL, Wattenmaker WD, Hampson SE (1987) Family resemblance, conceptual cohesiveness, and category construction. Cognitive psychology 19(2):242-279.

Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review 63(2):81.

Moe WW, Trusov M (2011) The value of social dynamics in online product ratings forums. Journal of Marketing Research 48(3):444-456.

Moreland, R. L. (1987) The formation of small groups. In C. Hendrick (Ed.), Group processes: Review of personality and social psychology (Vol. 8, pp. 80–110). Newbury Park, CA: Sage.

Pavitt C, Johnson KK (1999) An examination of the coherence of group discussions. Communication Research 26(3):303-321.

Peng H, Li J, Song Y, Liu Y (2017) Incrementally learning the hierarchical softmax function for neural language models. Proceedings of the AAAI Conference on Artificial Intelligence.

Pennebaker JW, Booth RJ, Francis ME (2007) LIWC2007: Linguistic inquiry and word count. Austin, Texas: liwc. net.

Peterson L, Peterson MJ (1959) Short-term retention of individual verbal items. Journal of experimental psychology 58(3):193.

Picciotto R (2005) The evaluation of policy coherence for development. Evaluation 11(3):311-330.

Rawaf S, Allen LN, Stigler FL, Kringos D, Quezada Yamamoto H, van Weel C, Coverage GFoUH, Care PH (2020) Lessons on the COVID-19 pandemic, for and by primary care professionals worldwide. European Journal of General Practice 26(1):129-133.

Sears DO, Freedman JL (1967) Selective exposure to information: A critical review. Public Opinion Quarterly 31(2):194-213.

Shen D, Yang Q, Sun J-T, Chen Z (2006) Thread detection in dynamic text message streams. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 35-42.

Stella M, Ferrara E, De Domenico M (2018) Bots increase exposure to negative and inflammatory content in online social systems. Proceedings of the National Academy of Sciences 115(49):12435-12440.

Streufert SC (1973) Effects of information relevance on decision making in complex environments. Memory & Cognition 1(3):224-228.

Stromer-Galley J, Martinson AM (2009) Coherence in political computer-mediated communication: analyzing topic relevance and drift in chat. Discourse & Communication 3(2):195-216.

Susarla A, Oh J-H, Tan Y (2012) Social networks and the diffusion of user-generated content: Evidence from YouTube. Information Systems Research 23(1):23-41.

Swan K (2002) Building learning communities in online courses: The importance of interaction. Education, Communication & Information 2(1):23-49.

Sweller J (2011) Cognitive load theory. Psychology of learning and motivation, vol. 55 (Elsevier), 37-76.

Te'eni D (2006) The language-action perspective as a basis for communication support systems. Communications of the ACM 49(5):65-70.

Thomson R (2006) The effect of topic of discussion on gendered language in computer-mediated communication discussion. Journal of Language and Social Psychology 25(2):167-178.

Thongtan T, Phienthrakul T (2019) Sentiment classification using document embeddings trained with cosine similarity. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 407-414.

Tiao GC, Xu D (1993) Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. Biometrika 80(3):623-641.

Toubia O, Stephen AT (2013) Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter? Marketing Science 32(3):368-392.

Tumuluru AK, Lo C-k, Wu D (2012) Accuracy and robustness in measuring the lexical similarity of semantic role fillers for automatic semantic MT evaluation. Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, 574-581.

Villarroel Ordenes F, Ludwig S, De Ruyter K, Grewal D, Wetzels M (2017) Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media. Journal of Consumer Research 43(6):875-894.

Wan X, Yang J (2007) Learning information diffusion process on the web. Proceedings of the 16th international conference on World Wide Web, 1173-1174.

Wang C, Zhang X, Hann I-H (2018) Socially nudged: A quasi-experimental study of friends' social influence in online product ratings. Information Systems Research 29(3):641-655.

Wang M, Lu S, Zhu D, Lin J, Wang Z (2018) A high-speed and low-complexity architecture for softmax function in deep learning. 2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS) (IEEE), 223-226.

Wang Q, Kayande U, Jap S (2010) The seeds of dissolution: Discrepancy and incoherence in buyer–supplier exchange. Marketing Science 29(6):1109-1124.

Wang Y, Chaudhry A (2018) When and how managers' responses to online reviews affect subsequent reviews. Journal of Marketing Research 55(2):163-177.

Weger Jr H, Aakhus M (2003) Arguing in Internet chat rooms: Argumentative adaptations to chat room design and some consequences for public deliberation at a distance. Argumentation and advocacy 40(1):23-38.

Wise AF, Marbouti F, Hsiao Y-T, Hausknecht S (2012) A survey of factors contributing to learners'"listening" behaviors in asynchronous online discussions. Journal of Educational Computing Research 47(4):461-480.

Woerner SL, Yates J, Orlikowski WJ (2007) Conversational coherence in instant messaging and getting work done. 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07) (IEEE), 77-77.

Xiao B, Benbasat I (2015) Designing warning messages for detecting biased online product recommendations: An empirical investigation. Information Systems Research 26(4):793-811.

Yang J, Counts S (2010) Predicting the speed, scale, and range of information diffusion in twitter. Proceedings of the International AAAI Conference on Web and Social Media.

Zare F, Guillaume JH, Jakeman AJ, Torabi O (2020) Reflective communication to improve problem-solving pathways: Key issues illustrated for an integrated environmental modelling case study. Environmental Modelling & Software 126:104645.

Zhang L, Peng T-Q (2015) Breadth, depth, and speed: diffusion of advertising messages on microblogging sites. Internet Research 25(3): 453-470.