Understanding Influential Comments in Online Conversations

Jiwon Kang
Hanyang University
Republic of Korea
jiwonkang@hanayang.ac.kr

Daejin Choi
Seoul National University
Republic of Korea
djchoi@mmlab.snu.ac.kr

Eunil Park
Sungkyunkwan University
Republic of Korea
eunilpark@skku.edu

Jinyoung Han
Hanyang University
Republic of Korea
jinyounghan@hanyang.ac.kr

CSCW'18 Companion, November 3–7, 2018, Jersey City, NJ, USA

ACM ISBN 978-1-4503-6018-0/18/11.
https://doi.org/10.1145/3272973.3274054

Abstract
Online communication has become a common social activity in our daily lives. This paper investigates the roles of 'influential comments' that affect other comments in online communication. To this end, we collect and analyze a large-scale communication data from Reddit, which consists of 81 K news-related posts and their 3 M associated comments written by 400 K users. Using the collected data, we investigate how the influential comments affect the follow-up comments in a communication in terms of (i) topic similarity and (ii) revealed sentiment. Our work reveals that influential comments tend to more affect their follow-up descendant comments than the post in terms of topic similarity and revealed sentiment.

Author Keywords
Online Communication; Influential Comment; Reddit

Introduction
Online communication through social networking services, messengers, and message boards has become one of the popular and common social activity in our daily lives. With this trend, several studies have investigated how people communicate with each other in online communities such as Reddit. Choi *et. al* [2] analyzed online conversations in Reddit, and showed that the small number of users mainly generate the majority of the viral conversation. Liang [5] analyzed the conversations in the tech-related Q&A sub-

Figure 1: An example of a threaded conversation in the subreddit "r/world news".

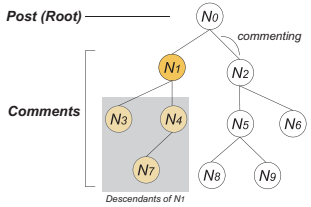

Figure 2: An example of a conversation tree with 9 comments. The child nodes of N_1 are N_3 and N_4. The descendants of N_1 includes N_3, N_4, and N_7.

reddit in Reddit, and revealed that posts, where active commenters participate, are likely to have high scores.

While those work has revealed valuable insights into understanding the patterns of online conversations, however, relatively little has been empirically known about whether and how specific comments (or user responses to a topic) affect other users' responses in active online conversations. For example, users in a conversation may follow a few comments' opinions or tones, which may result in a distorted public opinion to a particular alignment. Understanding the roles of such *'influential comments'*, which substantially affect other comments in a conversation, can provide important insights for journalists, policy makers, social scientists, or politicians who want to understand public opinion or voice of the crowd in a society.

To bridge such a gap, this paper seeks answers to the following research questions: *How can influential comments be identified in a conversation? How do the identified influential comments affect the follow-up comments in a conversation?* To address these questions, we investigate online conversations in Reddit, one of the popular online communities where people communicate with each other in a threaded conversation fashion (See Figure 1 as an example). Note that a threaded conversation consists of a post and its comments; a person brings up a subject by uploading a post, and other people can write a comment to the post, or to others' comments. We collect a large-scale conversation data from six news-related subreddits in Reddit. Using the collected data, we identify the influential comments, and investigate how the identified influential comments affect the follow-up comments in a conversation in terms of (i) topic similarity and (ii) revealed sentiment.

Materials and Method

In this section, we first present our dataset used in this paper. We then describe how to model a threaded conversation as a tree in Reddit.

Dataset

We analyze conversation data from the following news-related subteddits: *r/news*, *r/uncensored news*, *r/world news*, *r/uplifting news*, *r/true news*, and *r/fake news*. To this end, we collected the posts and their associated comments in those subreddits using the *Python Reddit API Wrapper* package (https://praw.readthedocs.io). The dataset includes 81,663 posts and their 3,300,415 associated comments written by 406,213 users from January 2017 to July 2017.

To investigate how influential comments affect the follow-up comments in an active conversation, we focus on the conversations that have more than 99 comments. As bots in certain subreddits, e.g., *r/world news*, often generate the first comments for automatically summarizing the news, we filter out those first bot-generated comments. As a result, we analyze 8,424 posts and their 2,610,610 comments written by 368,598 users.

Conversation Model

To model a threaded conversation, we define the notion of a *conversation tree* as an undirected tree with a post as a root node, $T = (N, E)$ where N is the set of nodes including the post and all the follow-up comments, and E is the set of edges, each of which links two nodes by commenting. Figure 2 exemplifies a conversation tree with a post, N_0, and its nine comments.

Identification of Influential Comments

To identify the influential comments that potentially affect other users' comments in a conversation, we consider the following comment node properties.

- **Comment Score**: The score of a comment is calculated as the number of upvotes minus the number of

	Score	Child	BC
Score	-	-	-
Child	0.52	-	-
BC	0.57	0.93	-

Note: Score = Comment Score; Child = Child Count; BC = Betweenness Centrality.

Table 1: Rank Correlation among the three node properties. All the *p*-values are smaller than 0.05.

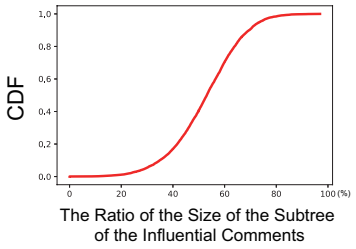

Figure 3: Ratio of the size of the subtree of the identified influential comments. Note that the average ratio is 0.52.

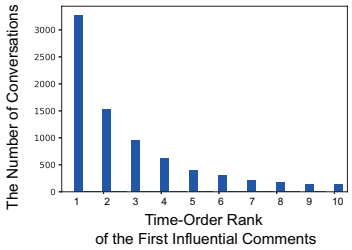

Figure 4: Histogram of the time-order rank of the first influential comments.

downvotes, which reveals how people agree with the given comment. A comment with high score tends to be displayed in the top of the comment section.

- **Child Count**: The child count is the number of child nodes in a conversation tree. We conjecture that the child nodes usually are affected by their parent node since they are directly replying to their parent node.
- **Betweenness Centrality**: The betweenness centrality [3] of a node is the number of the shortest paths that pass through the node. A comment with high betweenness centrality tends to (i) bridge other comments and/or (ii) have a large subtree.

To examine how the above three properties are related, we calculate the Spearman's rank correlation [4] among the properties, ranging between -1 and 1. As shown in Table 1, the correlation between the betweenness centrality and child count is 0.93, which means those properties are strongly correlated. In other words, a node with high betweenness centrality tends to have many child nodes. Since the rank orders by those two properties are almost identical, we only consider the betweenness centrality hereafter. Note that the comment score is also partially correlated with the child count as well as betweenness centrality; the correlation coefficients for the child count and betweenness centrality are 0.52 and 0.57, respectively. It implies that a comment with a high score may or may not have many child nodes or high betweenness centrality.

Finally, we identify influential comments based on both (i) comment score and (ii) betweenness centrality. That is, we make the two top 10 node lists in terms of those two properties in a conversation tree, and then find the overlapped nodes between two lists, which are regarded as influential comments in our study. On average, the number of identified influential comments in a conversation tree is 5.08.

Results

Characteristics of Identified Influential Comments

We first measure how identified influential comments contribute to the size of the corresponding conversation tree. To this end, we calculate the ratio of the size of the subtree of the influential comments to the size of the conversation tree. For example, in Figure 2, if N_1 is identified as an influential comment, the ratio of its subtree is 0.4. As shown in Figure 3, the influential comments in many conversation trees tend to generate a large number of follow-up comments. For example, in more than 25% of the conversation trees, the influential comments are responsible for more than 60% of nodes in their corresponding conversation trees. It implies that the influential comments substantially contribute to the total size of the conversation tree.

We then analyze when the influential comments are written in a conversation tree. To this end, we find the first influential comment, which is written firstly among the identified influential comments in their conversation tree. We then plot the time-order of the first influential comment of each conversation tree in Figure 4. As shown in Figure 4, among the first influential comments, 38% of them are written firstly; 18% of them are written secondly in their conversations. It reveals that comments written in an early stage of a conversation tend to play influential roles in the conversation.

Roles of Influential Comments

To analyze how influential comments play roles in a conversation, we investigate how they are similar to their descendant comments in terms of topic similarity and sentiment. Our conjecture is that the influential comments can affect the opinion or sentiment of the follow-up discussions, i.e., descendants of the influential comments. Since a post usually drives its conversation, we compare the influential comments and the post in a conversation tree.

To measure the topic similarity between two documents,

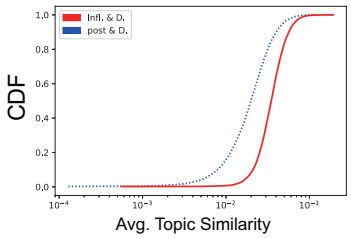

Figure 5: Average topic similarities between (i) influential comments and their descendants, and (ii) a post and the descendants of the influential comments. The Kolmogorov-Smirnov test shows that the distributions are statistically different (*p-value* < 0.001).

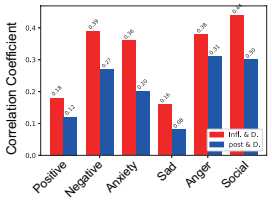

Figure 6: Pearson correlation coefficients of the sentiment similarities between (i) influential comments and their descendants, and (ii) a post and the descendants of the influential comments (*p-value* < 0.001).

i.e., a post and a comment, or two comments in our case, we use the Term Frequency Inverse Document Frequency (TF-IDF) [1] for quantifying the document relevancy between two documents. We first remove the stop-words and adopt the Porter Stemming provided by *Natural Language Toolkit* (http://www.nltk.org). We then apply the TF-IDF weights for given texts, and calculate the cosine similarity between two word vectors, ranging from 0 to 1. If the cosine similarity is 1, it means two texts are almost identical. Figure 5 shows the average topic similarity among (i) influential comments and their descendants, and (ii) a post and the descendants of the influential comments. As shown in Figure 5, topic similarities among influential comments and their descendants are much higher than those among a post and the descendants of the influential comments, implying that influential comments tend to more affect their descendant comments than the post.

We next investigate how the influential comments affect the sentiment of the follow-up discussions by using the LIWC (http://liwc.wpengine.com), a popular tool for sentiment analysis. We first calculate the positive, negative, sad, anger, anxiety, and social scores of a post and all the follow-up comments. We then calculate the average sentiment scores for the influential comments, and for the descendants of the influential comments. For measuring the sentiment similarity between two groups, i.e., influential comments and their descendants, we calculate the Pearson correlation coefficients between average revealed sentiment scores of the two groups.

As shown in Figure 6, the sentiment similarities between influential comments and their descendants are much higher than those between a post and the descendants of the influential comments. This means that the follow-up comments of the influential comments tend to reveal more similar sentiments with the influential comments than the post.

Concluding Remarks

This paper investigated how influential comments in a conversation affect the follow-up comments, in comparison with the post. We found that comments written in an early stage of a conversation tend to play influential roles. We also revealed that influential comments tend to more affect their descendant comments than the post in terms of (i) topic similarity and (ii) revealed sentiment. Our future work includes (i) investigating how troll comments affect other users' opinions and (ii) developing a model for predicting influential comments in the early stage of a conversation.

Acknowledgement

This work was supported in part by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A5B8058870). The corresponding author of this paper is Jinyoung Han (jinyoung-han@hanyang.ac.kr).

References

[1] Aizawa, A. An information-theoretic perspective of tf–idf measures. *Information Processing & Management 39*, 1 (2003), 45–65.

[2] Choi, D., Han, J., Chung, T., Ahn, Y.-Y., Chun, B.-G., and Kwon, T. T. Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In *Proc. of ACM COSN* (2015).

[3] Freeman, L. C. Centrality in social networks conceptual clarification. *Social networks 1*, 3 (1978), 215–239.

[4] Kokoska, S., and Zwillinger, D. *CRC Standard Probability and Statistics Tables and Formulae*. CRC Press, 1999.

[5] Liang, Y. Knowledge sharing in online discussion threads: What predicts the ratings? In *Proc. of ACM CSCW* (2017).