Understanding Characteristics of Catalyst Users in the WallStreetBets Community

Ehsan-Ul Haq[§], Yiming Zhu[§], Zijun Lin[¶], Haodi Weng[‖], Gareth Tyson[§], Lik-Hang Lee[†], Reza Hadi Mogavi[**], Tristan Braud[*], and Pan Hui[*§]
[*]The Hong Kong University of Science and Technology, HKSAR
[†]The Hong Kong Polytechnic University, HKSAR
[§]The Hong Kong University of Science and Technology, Guangzhou
[¶]London School of Economics and Political Science, UK, [‖]Imperial College London, UK
[**]University of Waterloo, Canada
Email: {euhaq,yzhucd}@connect.ust.hk {jezzalamji,wengvictor5}@gmail.com lik-hang.lee@polyu.edu.hk
rhadimog@uwaterloo.ca {gtyson,braudt,panhui}@ust.hk

***Abstract*—WallStreetBets (WSB), a Reddit community, impacted stock markets during the 2021 GameStop Short Squeeze. We examine the content and user properties that influence engagement in WSB. Despite WSB's association with emojis and informal terms, engagement among community members depends on more than surface-level factors. Although emojis are commonly used, they are not as effective at fostering interactions among users. Community members engage more with posts that have longer and topic-specific text. Simply producing a high volume of posts is not enough to attract an audience. Consistent topical focus, reciprocal interactions, and previous authorship of catalyst posts influence engagement. WSB posts, regardless of length, generally remain relevant to the community's theme of stock trading. Our findings provide insights into WSB engagement patterns and can be useful for downstream research, such as financial predictive tasks using WSB data.**

***Index Terms*—Reddit, Engagement, WallStreetBets, Catalyst**

I. Introduction

WallStreetBets (WSB) community on Reddit came to the fore in 2021 during the GameStop Short Squeeze,[1] where the community rallied against investors attempting to short GameStop stock [1]. Attracting worldwide attention, the community experienced sudden growth in new users and posts. We posit that this sudden growth may affect content reach and hence affect users' engagement with each other. Quantifying engagement predictors can help support effective and sustainable community growth. Thus, the WSB community offers an opportunity to explore the interaction and engagement characteristics of communities that experience a sudden influx of users. WSB is particularly important to study due to its relevance and value to financial markets, and is the 47th largest community on Reddit [2].

We present the first study investigating the factors influencing engagement within WSB during the 2021 GameStop Short Squeeze, with a comparative analysis of the periods before and after the short squeeze. WSB stands out due to its distinctive attributes, such as its unique writing style, utilization of emojis to represent market trends and informal terminology. A recent study on WSB revealed that new users extensively adopted this style during the short squeeze surge [1].

We analyze the characteristics of WSB users and content that drive engagement within the community. We examine two patterns of participation: catalyst and interacting. Catalyst engagement involves community members replying to comments on the main posts, whereas interacting refers to direct comments on the main post, with no replies to any comment. Catalyst engagement reveals multi-user interactions that are not captured by the cumulative number of comments. We consider WSB-specific factors like emojis, text length, topics, and user interactions. Our findings are as follows: (*i*) We characterize community participation based on the communication style and topical choices. Catalyst posts tend to be longer and have fewer emojis, which contradicts WSB norms. This applies to catalyst users as well, unaffected by the short squeeze. (*ii*) We quantify the impact of these features on WSB engagement. Earlier catalyst posts and reciprocal interactions from catalyst users increase the likelihood of posts becoming catalyst posts, in addition to the WSB communication norms and the authors' topical preferences.

II. Related Work

Characterisation of community members' interaction patterns and topical preferences is a widely studied field [3], [4]. The focus of such studies varies from studying user behaviors in niche-oriented communities [5] to inspecting large-scale platform adaptability and usage [6]. The engagement generated by a user is usually associated with the role that a user plays in a community [6], [7]. Traditional approaches to studying users with influential roles typically center around network-based

[1]https://en.wikipedia.org/wiki/GameStop_short_squeeze

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
http://dx.doi.org/10.1145/3625007.3627595

analysis. For example, researchers have identified users who strategically position themselves within follower networks to act as information brokers. Additionally, users and their content have been examined in terms of their authoritative role within the community, including the study of moderators [8]. Users are often studied based on their content creation and consumption patterns. For example, on Reddit, users are classified as initiators (prolific post creators), commentors (primarily commentors), or attractors (users who attract comments) [6]. Choi et al. also note the reciprocal nature of interactions among Reddit users. Catalyst users, on the other hand, generate discussions among friends on Facebook [7]. Previous research has primarily focused on network analysis; our study examines the content characteristics associated with users based on engagement information.

Online users employ various ways to exert their position and share content [9]. The linguistic characteristics and writing styles of users' associated roles can be used to study such user dynamics in the community [10]. The network-based characterisation of content can highlight the information flow, but language can offer insights into the authors' attention to topics and their opinions. The communication style is important, particularly when the communities are topic-oriented. Focused communities usually carry their linguistic norms, such as emojis in the case of WSB or gaming communities [11]. Stylometric properties are used in tasks such as text and authorship classification and vary from syntactic (text length) to lexical analysis (recurrent patterns, use of words) [12], [13]. Frequently used stylometric properties include the length and readability of the text [14], the use of a lexicon [15], and emoticons [16]. Our work focuses on linguistic and community feedback and reveals information about user-engagement preferences in WSB.

III. Methodology

Data Collection. Using Pushshift API [17], we collect data from January 2020 to June 2021, consisting of 1,553,189 posts from 568,756 users. These posts generate 39,039,301 comments with a tailed distribution of comments per post ($min = 0, max = 100002, \mu = 25.13, median = 0$, $\sigma = 775.39$). The short squeeze period in January contains 536,691 posts from 250,967 users. This covers 7,442,010 comments (comments per post: $min = 0, max = 97519, \mu = 13.87, median = 0$, $\sigma = 717.20$). From January, 395,053 (73.61%) posts and 2,352,173 comments (28.96%) were either deleted or removed. There are 746,659 users in the dataset. The number of newly subscribed users to WSB (who began sending posts and comments in January 2021 but not before) is 615,525, and the number of earlier subscribed users is 131,133. Excluding deleted users, there are 783,659 posts sent by new users and 304,357 posts from earlier users.

Defining Catalyst and Interacting Posts and Users. To systematically analyse communication, we divide users and posts into two core groups and a corresponding control group.

Catalyst Posts and Users : WSB posts where community members also reply to comments, in addition to the main post, are referred to as catalyst posts, as described by [7]. Users who create such main posts are referred to as catalyst users.

Interacting Posts and Users: WSB posts that have received some comments, but no reply to any comment are referred to as interacting posts, and their authors as interacting users. Note, catalyst posts are a subset of interacting posts. But not all interacting posts are catalyst posts. We treat these post types and their authors as mutually exclusive. In our dataset, only 28.11% posts have at least one comment. This also gives an upper bound on the number of catalyst posts. Suggesting both interacting and catalyst posts follow a heavy-tailed distribution.To account for this skewed distribution, we only look for the posts that can amass a large response from the community within each of the three time periods. Hence, we decide to only look at the posts with top-5% comments. For catalyst posts, we take the top 5% of posts with comments with at least one reply. For interacting posts, we take the top 5% of posts based on their cumulative number of comments. We get 81,025 and 11,392 posts for the catalyst and interacting categories, respectively.

Control Group Posts and Users: We find that 42,957 users generate catalyst posts, and 7,800 users generate the interacting posts. Hence, comparing a small sample of users and posts that receive comments with a large sample might skew the results [7]. Thus, we first extract the authors of the interacting and catalyst posts and calculate the mean number of posts per author. We then extract the remaining users who have posted at least the mean of other users but are not part of other user groups. This controls the post-volume when comparing the user groups. The posts of this set of users are called control posts, whereas this set of users is called control users. A total of 28,210 users are identified as control users.

Time Segmentation. To address the extreme deviation in community norms during the short squeeze, we split the data set into three periods: before the surge data (year 2020), during the surge data (January-March 2021), and after the surge (April-June 2021) data. Within each period, we compare three sets of authors and posts, giving us nine categories: (*i*) Catalyst Before, (*ii*) Catalyst During, (*iii*) Catalyst After. (*iv*) Interacting Before, (*v*) Interacting During, (*vi*) Interacting After, (*vii*) Control Before, (*viii*) Control During, and (*ix*) Control After.

IV. Analysing Catalyst Posts and Users

We examine the communication style of the WSB, including the use of emojis and text length, as well as the topical choices made by engaging posts and users. We analyze the distinctions between two categories: 1) catalyst posts and 2) catalyst users, compared to other post- and user types, respectively.

A. Comparing Catalyst Post

WSB-specific Linguistic Features of Posts. We focus on analyzing specific linguistic markers unique to WSB: the use of emojis and the length of the text, which can indicate the seriousness of the content [12], [14]. We measure post

length by counting white-spaced words after removing emojis and URLs. Due to differences in sample sizes and the non-normal distribution of the data, we employ the non-parametric Kruskal-Wallis (KW) test to quantify differences among post-types, followed by pairwise comparisons using Dunn's Test with a Bonferroni correction [18], [19].

We observe a significant difference in the writing styles among users. KW statistics for emoji use in the title of posts ($\chi^2(8) = 3988.3, p < 0.0001$), for emojis in the text ($\chi^2(8) = 3317.3, p < 0.0001$), length of title ($\chi^2(8) = 2880.5, p < 0.0001$), and length of text ($\chi^2(8) = 21942, p < 0.0001$) are statistically significant. That means that the observed user groups' posts have different writing styles based on observed variables. We then use Dunn's Test to compare the groups pair-wise.[2] We note that catalyst posts are the lengthiest, followed by interacting posts. In addition, catalyst posts use fewer emojis in the title than control posts in either subset of the data split. However, they have more emojis in the text; this could be related to the larger post size (more number of words) and hence more emojis. Fewer emojis in the title suggest that the title is more descriptive than the visual markers.

Topics of the Posts. Next, we examine the topic differences among posts within each category and period. We utilize Empath [20], a lexicon-based topical categorization method used for linguistic characterization [21]. Empath provides normalized word counts for various topic categories. Figure 1 illustrates the average topics per post (post text only) for the top 10 most prevalent topics in each post group and period, with error bars representing a 95% confidence interval.

Figures 1a, 1b, and 1c show that catalyst posts have a higher focus on the top 10 topics in the dataset as compared to other posts. We see that both catalyst and interacting posts have a higher focus on topics related to economics, money, and shopping (representing the buying and selling of stocks) across all periods. The difference in the average use of these topics is more than double between posts with and without engagement. It suggests, in general, that posts focusing on the main theme of the community get engagement. This is further evident since topic preference and usage remain similar in periods for the catalyst posts. When combining the observations from the previous sections, we note that catalyst posts have more words (suggesting more detailed posts), and focus more on topics related to buying and selling of stocks. Whereas the non-engaging posts are shorter and have relatively less focus on WSB thematic topics on per post basis. Although titles share similar topical usage across all types of posts, the post text is an important factor for engagement.

B. Comparing Catalyst Users

We now look at differences among three sets of users: catalyst, interacting, and control group.

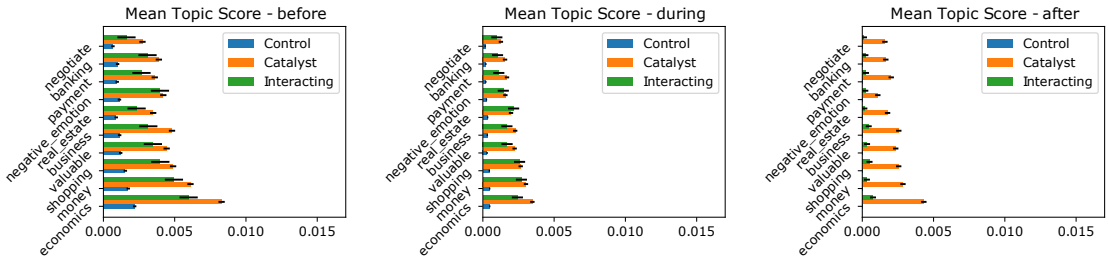
Linguistic features from users. We look at how detail-oriented users are in their writing styles and follow community-specific characteristics. We first filter the catalyst, interacting, and the control group authors and get their posts. We use the Kruskal-Wallis test for each of these posts, followed by a pair-wise comparison. Our results show a significant difference in the writing styles of different user groups. Kruskal-Wallis test among all the groups shows that emoji use in their posts' titles ($\chi^2(8) = 15625.4, p < 0.0001$), for emojis in the text ($\chi^2(8) = 662.16, p < 0.0001$), length of title ($\chi^2(8) = 1408.4, p < 0.0001$), and length of text ($\chi^2(8) = 5384.6, p < 0.0001$) are statistically significant.[2]We notice that the catalyst users generally have a longer title than the rest of the users (median of eight words *vs.* interacting users with a median of 6 words). The catalyst users also write longer posts compared to users in the control group. However, the catalyst users have shorter posts (median 57 words) during the surge compared to the before surge period (median 92 words). Interacting users have shorter posts than catalyst users in all periods. Our findings also align with [1], *i.e.,* that during January, communication becomes shorter than usual. However, in the post-surge period, the length of posts grew almost double to the surge period, with a median number of words in the posts being 254 and 198 for catalyst and interacting posts, respectively. The WSB communication after the short squeeze months has not come back to normal yet; however, patterns for engagement in WSB are relatively consistent throughout the three periods, showing higher engagement with longer posts and containing fewer emojis in the posts. This suggests that after the short squeeze period, there is more extensive participation from users in the community.

Topical choice of users. We next look at the users' choice of topics, following the same approach as in the previous section. We note that not all posts from catalyst users may generate engagement. Hence, topical analysis of all posts from these users throws light on a holistic view of their topics preferences. In Figure 2, we report the average use of topic-per-post for the top 10 topics in the posts' texts.

The topics discussed by the user groups align consistently with the community's topic preferences. By utilizing normalized values in our topical lexicons, we can determine whether a text is more focused on a particular topic based on higher normalized values, resulting in a higher average for that topic. The average values for the topics in Figures 2a and 2b show that interacting users stay more focused on these topics as compared to catalyst users, and Figure 2c depicts that catalyst users surpassing the average in after period. Top 10 topics within user groups, and topics for titles are reported in supplementary information.[2]

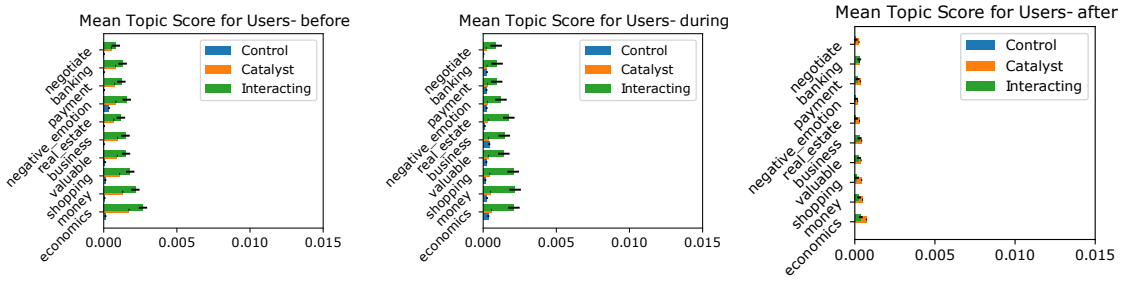
Catalyst users' specific topic analysis of posts shows different patterns compared to only catalyst posts topics. We see that catalyst posts have on average higher focus on topics related to buying and selling stocks as compared to interacting posts; however, in terms of overall posts from their respective authors, catalyst users have less focus on these topics as compared to interacting users. This suggests that while topics are an important factor for engagement, user-

[2]The results are available at https://drive.google.com/file/d/1YLilClwIXNOqDGVVuZmVPpscpvNTOLyL



(a) Text - Before (b) Text - During (c) Text - After

Fig. 1: Topic focus of three post types (Catalyst, Interacting, and Control) during three periods (Before, During, and After). Catalyst posts have a higher focus on the topic in text of the posts.



(a) Text - Before (b) Text - During (c) Text - After

Fig. 2: Topical focus of the posts from three user types (Catalyst, Interacting, and Control) during three periods (Before, During, and After).

specific characteristics must also help users for engagement.

V. Quantifying Community Engagement

We now quantify the effect of the above-distinguishing (topics and stylometric) features on WSB engagement. We also include other user-specific traits: their previous history of writing catalyst posts, preference to a topic, reciprocal interaction with other engaging users, and the time period. We formulate the task using Logistic Regression to predict whether a post will be a catalyst post. We also combine features with a period as an interaction term to deepen our understanding of community engagement over time. In total, we test four models: (*i*) A model only with four stylometric variables (count of words and emojis in title and text). (*ii*) A model with five user-specific features. This includes the user's previous history of writing catalyst posts (Prev. catalyst). A user's topic preference is the probability of writing on a specific topic from all the topics the user has written. We then include the community trend by including a variable if the most dominant topic of a post is a trending topic in the community within 24 hours (calendar day of the post). We also include whether the post author has reciprocal engagement with the commenters, *i.e.,* whether the author has previously commented on the commenters' posts. (*iii*) A model that combines all the variables of i) and ii). (*iv*) A model that includes the time-period variable as an interaction term to study the effect of all variables in three periods. The results show that both the writing styles and the users' preference for topics affect catalyst engagement.[3] In terms of *textual features*, more emojis in a post's title negatively impact the engagement. More specifically, the difference in the logs of the expected count of the post being a catalyst post is -0.057 when an emoji is added to the title. The post length has a positive effect (+.01x) on the engagement.

We note positive and marginally high estimates for *user-specific features*: reciprocity, previous catalysts, and topic preference compared to the textual variables. The topic preference of a user is the strongest predictor with odd difference of 39.098 in Model 4. Interestingly, it has higher odds during surge as compared to the before (2020) period, followed by the after period. This suggests that users, who generate catalyst engagement, have written many posts on the same topic and established their command of certain topics. Hence, community engages on such topics with the selected users.

The impact of previous catalyst posts is positive, albeit, lower (0.747) in the surge period and becomes higher after the surge (1.019) than the before period. This behavior can be related to the short squeeze resulting in a new user influx and content. This influx might have interrupted the regular catalyst users' communication. Alternatively, the posts did not get the exposure as before due to rapid content generation, resulting in lower estimates for this predictor variable. The estimate of this value increased in the post-surge period,

[3]All regression results are reported in supplementary information https://drive.google.com/file/d/1YLilClwIXNOqDGVVuZmVPpscpvNTOLyL

suggesting more consistent post patterns from catalyst users.

In summary, longer text and fewer emojis in post titles lead to higher engagement. This pattern hold across the time periods. Users' history of writing engaging and topic preferences strongly affect catalyst engagement. Reciprocity has a greater impact during the short squeeze period compared to before and after, while the influence of previous catalyst posts reduces during the short squeeze.

VI. Discussion and Conclusion

We analyzed users' engagement in the WSB community, with a focus on WSB-specific communication features across three time periods. We found that posts with longer text consistently engage more users, even during the short squeeze period, when posts are relatively shorter. Interestingly, the use of emojis in titles has a negative impact on catalyst engagement. This finding challenges the common perception that emojis, such as "to the moon" or "rocket" positively influence the community's stock market discussion. We assume that shorter messages and the use of emojis serve as short guidelines for specific actions, rather than initiating discussions or drawing attention to specific topics. For example, a highly engaging catalyst post titled "An In Depth Look Into Carnival Cruise Corp. (CCL) And Why This Stock Is A Ticking Time Bomb For Gains Come 2021" had 150 comments with at least one reply. Previous research has associated longer text length with the seriousness of messages [12], [14].

We observed a similar behavior among catalyst users, who tend to write longer posts and fewer emojis in the titles. Our control group selection for group comparisons indicates that mere activity does not guarantee engagement. Users need to choose and establish their preferred topics within the community. The community shows a preference for engaging with these users, making it challenging for newcomers to WSB to attract a larger user base. It's important to note that we did not qualitatively analyze the discourse of catalyst. Finally, Reciprocity positively impacts the catalyst engagement in WSB, similar to other platforms [22]. During the surge, reciprocity had a greater impact, while the influence of previous catalyst posts decreased in comparison to that of other periods. We speculate that this could be due to users primarily engaging in reciprocal interactions. These users may follow each other, making seeing and engaging with posts in their timelines easier. Exploring the role of social connections on engagement could be a potential direction for future research.

We acknowledge certain limitations and areas for future research. Our study is solely based on feedback-based measures (comments) and linguistic characteristics, as follower-based network information is unavailable on Reddit. Additionally, we do not account for time-fixed effects for individual users, instead examining their collective behavior. Our findings are specific to the WSB community, and there is potential to explore communication norms on other subreddits. Furthermore, our analysis only considers the cumulative number of replies to comments, which may overlook nuanced interactions such as temporal dynamics and sentiment. Future research will delve into detailed reply networks, utilizing a network-based approach to better understand the catalyst user's position within broader information cascades.

References

[1] E.-U. Haq, T. Braud, L.-H. Lee, A. K. Vallapuram, Y. Yu, G. Tyson, and P. Hui, "Short, colorful, and irreverent! a comparative analysis of new users on wallstreetbets during the gamestop short-squeeze," in *In Companion Proceedings of the WebConf*, 2022.

[2] Reddit, "Reddit - Dive into anything — reddit.com." https://www.reddit.com/best/communities/1/. [Accessed 10-10-2023].

[3] N. Ali-Hasan and L. A. Adamic, "Expressing social relationships on the blog through links and comments.," in *ICWSM*, 2007.

[4] T. Hu, J. Luo, and W. Liu, "Life in the" matrix": Human mobility patterns in the cyber space," in *ICWSM*, 2018.

[5] J. Han, D. Choi, J. Joo, and C.-N. Chuah, "Predicting popular and viral image cascades in pinterest," in *ICWSM*, 2017.

[6] C. et al, "Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors," in *ACM COSN*, 2015.

[7] M. Saveski, F. Kooti, S. Morelli Vitousek, C. Diuk, B. Bartlett, and L. A. Adamic, "Social catalysts: Characterizing people who spark conversations among others," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–20, 2021.

[8] D. Wadden, T. August, Q. Li, and T. Althoff, "The effect of moderation on online mental health conversations," *ICWSM*, 2021.

[9] Y. Liang, "Knowledge sharing in online discussion threads: What predicts the ratings?," in *Proceedings of the ACM CSCW*, 2017.

[10] S. Albota, "Linguistically manipulative disputable semantic nature of the community reddit feed post," in *CEUR Workshop Proceedings*, 2021.

[11] S. L. Graham, "A wink and a nod: The role of emojis in forming digital communities," *Multilingua*, vol. 38, no. 4, pp. 377–400, 2019.

[12] K. Lagutina, N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, I. Paramonov, and P. Demidov, "A survey on stylometric text features," in *FRUCT*, IEEE, 2019.

[13] M. Eder, M. Piasecki, and T. Walkowiak, "An open stylometric system based on multilevel text analysis," *Cognitive Studies*, no. 17, 2017.

[14] A. Tsou, "How does the front page of the internet behave? readability, emoticon use, and links on reddit," *First Monday*, 2016.

[15] M. Manabe, K. Liew, S. Yada, S. Wakamiya, E. Aramaki, *et al.*, "Estimation of psychological distress in japanese youth through narrative writing: Text-based stylometric and sentiment analyses," *JMIR Formative Research*, vol. 5, no. 8, p. e29500, 2021.

[16] M. De Choudhury and S. De, "Mental health discourse on reddit: Self-disclosure, social support, and anonymity," in *AAAI ICWSM*, 2014.

[17] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," in *AAAI ICWSM*, 2020.

[18] A. Dinno, "Nonparametric pairwise multiple comparisons in independent groups using dunn's test," *The Stata Journal*, vol. 15, 2015.

[19] E.-U. Haq, L.-H. Lee, G. Tyson, R. H. Mogavi, T. Braud, and P. Hui, "Exploring mental health communications among instagram coaches," in *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 218–225, IEEE, 2022.

[20] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *ACM CHI*, 2016.

[21] N. A. Ghani, S. Hamid, I. A. T. Hashem, and E. Ahmed, "Social media big data analytics: A survey," *Computers in Human Behavior*, 2019.

[22] G. Comarela, M. Crovella, V. Almeida, and F. Benevenuto, "Understanding factors that affect response rates in twitter," in *ACM HT*, 2012.