Leveraging Twitter data to analyze the Virality of Covid-19 tweets: A Text Mining Approach

Krishnadas Nanath* and Geethu Joy

Middlesex University Dubai

(author last version- accepted in ***Behaviour & Information Technology*** journal)

ABSTRACT

As the novel coronavirus spreads across the world, work, pleasure, entertainment, social interactions, and meetings have shifted online. The conversations on social media have spiked, and given the uncertainties and new policies, COVID-19 remains the trending topic on all such platforms, including Twitter. This research explores the factors that affect COVID-19 content-sharing by Twitter users. The analysis was conducted using 57,000 plus tweets that mentioned COVID-19 and related keywords. The tweets were subjected to the Natural Language Processing (NLP) techniques like Topic modeling, Named Entity-Relationship, Emotion & Sentiment analysis, and Linguistic feature extraction. These methods generated features that could help explain the retweet count of the tweets. The results indicate that tweets with named entities (person, organization, and location), expression of negative emotions (anger, disgust, fear, and sadness), reference to mental health, optimistic content, and greater length have higher chances of being shared (retweeted). On the other hand, tweets with more hashtags and user mentions are less likely to be shared.

Keywords: Covid-19; retweet; content sharing; mental health; emotion-focused content; information-focused content; social media.

1. Introduction

The COVID-19 pandemic impact is evident in all people that were affected worldwide. To reduce the spread of the virus, government authorities have implemented preventive measures, which have apart from slowing spread also restricted human movement and activity. With social distancing becoming the new norm, activities have shifted online, including both work and education. Besides, there has been an exponential rise in ecommerce. With the growth of users using online platforms to communicate and work, the conversations around COVID-19 continue to expand on social media platforms (Abbas et al., 2020). Users seek information from platforms like Twitter to understand the global sentiments around the pandemic.

Twitter has long been used by government entities and organizations to understand trends and disseminate information. COVID-19, being the first pandemic in the current digital age, has initiated conversations around the pandemic, with millions of users participating in discussions on social media platforms. On 11th March, when Covid-19 was declared a pandemic by the WHO, the number of social media COVID-19 mentions crossed 19 million, according to the analytics platform, Sprinklr. In early April, a report by Twitter stated that COVID19 tweets were being retweeted every 45 milliseconds. However, given the scale of tweets, disproportionately, only a small fraction of tweets go viral and reach a sizable audience (Adamic et al., 2000). Hence, it could be interesting to explore the factors that affect the resharing behaviors, especially in a pandemic situation like COVID-19.

Scholars in the past have taken a keen interest in understanding the salient features of Twitter to predict the actual reposting behavior (Li et al., 2018; Lin et al., 2016b). Several factors have been explored

* Krishnadas Nanath (Corresponding author): username.krishna@gmail.com
Word Count: 10, 371

in the context of content sharing - representational features of Tweets (Son et al., 2019), message features (Bruns et al., 2012), user characteristics (Hughes et al., 2014), and source structural factors (Wang and Zhuang, 2017). These factors help to explain the content-sharing behavior of users on Twitter and social media. On the other hand, the intersection of COVID-19 and Twitter has also attracted several scholars to explore the domain in 2020. While a few studies have conducted a content analysis of COVID-19 Twitter posts (Ahmed et al., 2020; Park et al., 2020), others have looked at COVID-19 information and misinformation dissemination on Twitter (Rosenberg et al., 2020; Rufai and Bunce, 2020).

While several studies have studied information sharing behavior on Twitter, even with specific reference to the COVID 19 pandemic, however, there is scope to glean deeper insights into the subject. There are several research gaps that are noteworthy and offer strong motivations for this research. First, language dimensions have essential effects on sharing behavior. While studies have explored the importance of language characteristics on resharing behavior in the context of online ads, their contribution to a pandemic situation on social media remains unclear. Second, although studies have looked at virality in the context of social crisis (Xu and Zhang, 2018), they have not adequately examined the role of emotions, content features, and linguistic styles on information sharing during a pandemic.

The objective of this study is to examine the role of emotion-focused content, information-focused content, linguistic features, and content features (topics) on the COVID-19 information-sharing behavior by social media users. Three research questions are addressed in this study:

- Does emotion-focused content play a role in the sharing of COVID-19 content on Twitter?
- Does information-focused content play a role in the sharing of COVID-19 content on Twitter?
- Do the linguistic features of a tweet play a role in the sharing of COVID-19 content on Twitter?

This study aims to reduce manual coding by using machine-learning features like Natural Language Processing, Topic modeling, Named Entity-Relationship, and Sentiment analysis. The article is organized as follows: Section 2 provides the literature review and conceptual model; Section 3 is the research method section, which describes the machine-learning techniques and modeling details; Section 4 presents the results of the model; and finally, the article concludes with discussion, implications, and future research directions.

2. Literature Review

With the advent of the internet and related technologies, social media has become a ubiquitous platform for people to network and share content online. Several studies have explored the speed and extent to which information is shared on globally networked social media and its ease of use, and how this is transforming public discourse worldwide. Asur and Huberman (2010) studied the prediction power of information dissemination on social media on real-world outcomes. Mayrhofer et al. (2020) found that in comparison to disclosed advertisements and brand posts on Facebook, user-generated content leads to higher purchase intention. This explains the power of understanding user-generated content on social media platforms from the impact, reach, and resharing perspective.

Twitter has been a popular platform to explore information dissemination research in literature. Wang and Yang (2020) performed a content analysis of various tweets and concluded that the organization's dialogic communication greatly influenced their public engagement. A larger-scale analysis (52.5 million tweets) was conducted by Roy et al. (2020) to evaluate social media's effectiveness in communicating necessary information to vulnerable communities during calamities or emergencies. Several studies have used the analysis of tweets in the context of politics. Lee et al. (2020) studied the influence of politicians' Twitter blunders and citizens' responses to these tweets. Another research conducted in the context of Twitter and politics conducted by Bright et al. (2020) concluded that using Twitter as a broadcasting mechanism does not help to gain an edge in electoral outcomes.

Several studies in the past have explored the factors that influence social media content's virality. The large-scale and rapid diffusion of information online is often measured by using content-based and creator-based research methods. Owing to social media's ubiquitous nature, understanding what goes viral has received a great amount of scrutiny in recent years. According to Han et al. (2020), conceptual richness and predictive ability can be improved by adding content-creator interactions to the nomological network of virality (based on tweets). Skaza and Blais (2017) studied the role of hashtags in determining the virality of a tweet . Tellis et al. (2020) aimed to understand the factors that determine the virality of YouTube videos. They found emotions, brand prominence, and information content as significant contributors towards virality. The authors concluded that ads demonstrating positive emotions (amusement, inspiration, and warmth) have a higher tendency of being shared compared to the negative ones. Mendoza et al. (2010) found that conducting aggregate analysis on tweets can help identify reliable news from rumors.

It is interesting to observe that there are a number of studies that have explored the intersection of the COVID-19 pandemic and social media. The ubiquitous nature of social media ensured that COVID -19 soon spiraled into an infodemic (Zarocostas, 2020). Hashtags like #Coronovirus became the top trending hashtags on social media in 2020. By influencing and fragmenting social response (Kim et al., 2019), the online platform has now fastened up the epidemic process. Several studies have explored the concept of misinformation and fake news related to the pandemic on social media. Pennycook et al. (2020) identified that nudging the users to evaluate news accuracy before sharing it can increase the truth discernment in them. Similarly, several other studies have discussed the ill effects of social media during the COVID outbreak. Gao et al. (2020) state that the government must identify measures to combat the infodemic while combating the citizens' mental health problems, such as depression and anxiety. Zhou et al. (2021) proposed a depression classification model that uses multimodal features and TF-IDF to extract depression cues and polarities from emotion, topic, and domain-specific perspectives. Their results show an increase in depression after the COVID -19 outbreak. Allington et al. (2020) studied social media's role as a disseminator of misinformation and conspiracy theories during pandemics.

Depoux et al. (2020) brought to light how the fear of pandemic itself spread faster than the virus (Larson, 2018; McCauley et al., 2013). They found that social media became a conduit for spreading misinformation and panic among its users, spreading faster than even the virus (Wilson and Chen, 2020). Yang et al. (2020) Studied the fear generated by COVID-19 myths and divided them into five categories: the spread of infection, preventive measures, detection measures, treatment, and miscellaneous. Interestingly, fear was the strongest emotion that all categories of myths invoked in people. The fine-grained analysis of sentiment dynamics (Zhou et al., 2021) in LGA (local government areas) revealed that sentiment changed from positive to negative, despite it being positive during the initial days. The sentiment varied differently at different stages, especially towards the various policies introduced by the government.

The above discussion highlights the need to manage the perils of misinformation being spread across online platforms; however, it also provides a perspective for brands to identify content related to the pandemic and position their posts accordingly. Merchant and Lurie (2020) recommended that all online platforms should be integrated to control the developing pandemic situation and remodel aspects of readiness and precautionary actions for the future. The authors highlighted the crucial role that social media can play during the pandemic. For instance, the Facebook news feed now works by redirecting users to either the WHO or local health authority websites. Yin et al. (2020) proposed a framework that could give insights based on topic-level dynamics. Based on an analysis of 13 million tweets, the authors found that positive sentiment had a higher ratio than negative sentiment. Different topics from the topic-level analysis showed different polarities. Topics such as “stay home stay safe” had positive polarity, whereas some topics like “people death’’ consistently had negative polarity. A summary of key papers, their category, and methods is provided in Table 1.

Table 1. Summary of Literature Review

<table>
<tr><th>Category</th><th>Paper</th><th>Description</th><th>Method</th></tr>
<tr><td rowspan="3">Social Media Content analysis</td><td rowspan="3">Asur and Huberman (2010); Wang and Yang (2020); Roy et al. (2020)</td><td>Recommendation and filtering system for Instagram users to categorize and analyze images</td><td>SVM with Cloud API</td></tr>
<tr><td>The study examines how organizations can used social media to establish dialogic relationship with the public</td><td>Relative frequencies and Cohen's kappa scores of dialogic principal indices.</td></tr>
<tr><td>Studies the effectiveness of social media in spreading awareness during calamity</td><td>Social-Mediated Crisis Communication Model (SMCC)</td></tr>
<tr><td rowspan="2">Tweet analysis</td><td rowspan="2">Lee et al. (2020); Bright et al. (2020)</td><td>Investigates how individual beliefs are affected by politicians' Twitter blunders, and their reactions to politicians’ tweets.</td><td>Moderated mediation analysis</td></tr>
<tr><td>Studies the effectiveness of political campaign on social media using Twitter data</td><td>A pooled time series model and a first difference model</td></tr>
<tr><td rowspan="2">Virality in social media</td><td rowspan="2">Skaza and Blais (2017);Tellis et al. (2020)</td><td>Uses time series data to quantify the proliferation and popularity of trending hashtags</td><td>SIR model</td></tr>
<tr><td>Study concludes that emotions, brand prominence and information content affect the virality of YouTube videos</td><td>mixed-effects model</td></tr>
<tr><td>Research Notes on COVID information dissemination on social media</td><td>Zarocostas (2020); Depoux et al. (2020); Larson (2018); Merchant and Lurie (2020);Wilson and Chen, (2020)</td><td>Discusses how social media caused the COVID pandemic to spiral into an infodemic spreading misinformation and fear. The research papers also suggest how social media could be used to direct people to genuine information (with evidence)</td><td>NA</td></tr>
<tr><td>Social media and its impact on epidemic</td><td>Kim et al., 2019</td><td>Focuses on understanding the effect of media on epidemic forecast and</td><td>SIR model</td></tr>
</table>

		uncovering measures to slow down the spread	
COVID misinformation on Social media	Pennycook et al. (2020)	Discusses how nudging users to evaluate news accuracy before sharing helps increase truth discernment	Regression
Mental health	Gao et al. (2020)	Studies mental health problems during epidemic from exposure to social media	Multivariable logistic regressions
Social media Information dissemination	Allington et al. (2020)	Discusses how social media can influence information dissemination and the relationship between COVID-19 health-protective behaviors and COVID -19 conspiracy beliefs	Regression
Pandemic and public response	McCauley et al. (2013).	Discusses how pandemic can induce individual and social stress.	Qualitative analysis
Covid-19 Myths on Twitter	Yang.S et al (2020)	Understanding different categories of COVID 19 myths on Twitter and analyzing people's reaction to these myths	SRI epidemic model
COVID -19 tweets and depression dynamics	Zhou et al. (2021)	Depression cues and polarity from COVID 19 tweets	TFIDF with multimodal features from tweets
Covid19 sentiment dynamics	Yin.H et al., (2020)	Analyzes the sentiment dynamics from topics	Topic modelling

3. Conceptual Framework

Sharing online content is an integral part of modern life. Several studies have explored the reasons why people share content online (Boyd et al., 2010; Hopp et al., 2020; Lee et al., 2015; Oliveira et al., 2020). Sharing content on social media makes the content more visible and potentially goes viral. While some studies have attributed the virality in social media to be random (Cashmore, 2009), others have explored several dimensions of content to understand why it goes viral (Berger and Milkman, 2012). One of the intuitive reasons why people share content on social media is that they find it to be useful and feel that it could help others (Wojnicki and Godes, 2008). Some authors also view social content sharing as a social exchange value that could generate mutual benefit for the users who share and consume content online (Fehr et al., 1998). Others have argued that emotional aspects play a significant role in determining whether content will be shared or not (Heath et al., 2001). The chances of sharing emotionally charged content are higher as people sense their experiences or deepen their social connections (Peters and Kashima, 2007).

Twitter is one of the prominent components of social media and has gained attention due to its significant information diffusion potential (Savage 2011; Shirky, 2011). The retweet feature of this microblogging website allows users over the network to share content. While some reasons why users retweet are similar to those for the content sharing behavior on social media in general (Cha et al., 2010), there are explanations particular to Twitter's context. According to a few studies, tweet features such as URL and hashtag inclusion result in more retweets (Suh et al., 2010). Several other reasons for retweeting have been explored in the literature, and these include: initiating a conversation (Boyd et al., 2010), updating information (Hwang and Shim, 2010), self-explanation (Lee et al., 2012), maintenance of social relationships (Park and Jeong, 2011) and others. While these reasons summarize the intentions to share content on social media, it would be interesting to observe the factors that impact the retweets in a new context like COVID-19.

Content sharing on social media has seen different trends in the context of the ongoing pandemic (Nabity-Grover et al., 2020). Pre-COVID-19, sharing updates on travel, tourism, reviews of restaurants and hotels were common and received well. However, similar updates are most likely to receive fewer likes and retweets in the current situation and be perceived as a selfish act (Brown, 2020; Harris, 2020). When people retweet a post on Twitter, they advocate the content or want their followers to view it. Therefore, content that could receive negative evaluation and backlash in the pandemic context is usually avoided as a retweet. The change in sharing content user behavior on social media post-COVID-19 is classified into outside-in and inside-out (Nabity-Grover et al., 2020). Outside-in refers to the content that was very exciting for users to share but is frowned upon during the current COVID-19 pandemic. Sharing of such content is now seen as taboo, and users retweeting this could be the target of vitriol (Fox 2020; St. Michel, 2020). Inside-out, on the other hand, refers to previously unshared content that is socially encouraged post-pandemic. For example, health-related topics on Twitter, usually considered private and sensitive (Lin et al., 2016a), are now shared more freely for public welfare. Therefore, it would be interesting to understand the factors that affect the retweet (or sharing) of tweets not only in the times of COVID-19 but also those related to the pandemic.

3.1 *Role of emotion-focused posts in COVID-19 content sharing*

Content shared by users online can display various emotions, and these could either be positive or negative. The emotions generated by online content could lead to a variety of outcomes ranging from developing attitude towards the content generator (Edell and Burke, 1987) to recall (Stayman and Batra, 1991) and viewing time (Teixeira et al., 2012). In the context of social media and Twitter, the range of emotions could decide upon the sharing of the content (Akpinar and Berger, 2017) and retweets of the Twitter posts (Stieglitz and Dang-Xuan, 2013). Several studies have assessed the impact of emotionally driven content and found that it has more impact than information-driven content (Lee and Hong, 2016). Akpinar and Berger (2017) assessed the sharing of emotionally driven content and found a positive influence.

The first point of assessment in the context of emotion-focused posts on social media and content sharing is whether it is positive or negative. There is a general belief that positive content has a higher chance of being reshared compared to negative emotions (Godes et al., 2005). From a viral content theory perspective, Berger and Milkman (2012) also hypothesized the virality of positive content and found that users often share content for self-presentation purposes (Wojnicki and Godes, 2008), and retweeting positive content may reflect positively on them. Several studies have confirmed the influence of positive valence on content sharing (Hagerstrom et al., 2014; Nelson-Field et al., 2013). The negative emotions can include anger, disgust, fear, and sadness; positive emotions, on the other hand, can comprise of joy, trust, and surprise (Berger and Milkman, 2012).

Apart from the valence of these emotions (positive and negative), the physiological arousal or activation they evoke could also influence content (Smith and Ellsworth, 1985). These emotion arousal abilities also play an essential role in social transmission (Berger, 2011). Literature has suggested that emotion arousal could lead to potential actions like assisting others (Gaertner and Dovidio, 1977) or responding to offers (Brooks and Schweitzer, 2011). Retweeting tweets and sharing content requires action, and hence similar effects can be observed because of emotions. Therefore, a pack of emotions (anger, disgust, sadness, amusement, awe, inspiration, surprise, joy, and fear) evoked by social media content has been explored in the context of content sharing (Dafonte- Gómez, 2015; Hsieh et al., 2012; Nikolinakou and King, 2018).

From the context of the COVID-19 pandemic, examining the interplay between emotional valence/arousal on content sharing could be fascinating. While both valence and arousal could make emotions a vital aspect of content sharing, the type of emotions could differ in the context of the COVID-19 pandemic. This study takes up the inside-out perspective described by Nabity-Grover et al. (2020). These perspectives deal with topics that were considered harmful or were not shared earlier but are now socially encouraged. The topics could be related to health (COVID-19 context), mental health during the pandemic (Pfefferbaum and North, 2020), anxiety and depression (Ni et al., 2020), and others. Such topics from the inside-out theory perspective indicate that they could widely be shared and retweeted during the pandemic even if they stood no chance of a more significant impact pre-COVID-19 (Kordzadeh and Warren, 2017; Lin et al., 2016a). The reason for this could be attributed to the global difficulties that people go through during the pandemic. These difficulties could lead to a greater emotional connect between the content and the user. The impact of users' emotional connection with the content on social media on the sharing and virality of the content has been explored in the literature (Dobele et al., 2007; Phelps et al., 2004). A critical perspective that explains this emotional connection phenomenon is the theory of resonance (McDonnell et al., 2017). It conveys that some content has an advantage over others because it fits (or resonates) with the views and emotions of the audience (Schudson, 1989). This resonance in the form of retweets could further shape the media discourse (Ferree et al. 2002; Gamson and Modigliani 1989) in this context.

Therefore, this research argues that emotions (both positive and negative) could be influential in COVID-19 content sharing on Twitter. There are eight emotions used in this research, and therefore eight hypotheses are used to address the first research question - Does emotion-focused content play a role in the sharing of COVID-19 content on Twitter?

H1a-H1h: *Tweets evoking Anger (a), Anticipation (b), Disgust (c), Fear (d), Joy (e), Sadness (f), Surprise (g), and Trust(h) will have more retweets in the context of COVID-19 content.*

3.2 *Role of information-focused content in COVID-19 content sharing*

Information-focused content is typically verbally rich and often refers to the relevance and timeliness of the information (DeLone and McLean, 1992; Wand and Wang, 1996). It mostly involves delivering factual or evidential information. The content posted or developed in an information-focused style will typically involve factual descriptions about people (names), locations, organizations, and events (Tellis et al., 2019). The literature on communication suggests that information-focused content can influence individual attitudes toward information sharing (Chang et al., 2015; Maheswaran and Meyers-Levy, 1990). The factual content in the social media posts influences the decision of its impact. The extent to which the informative words are used is shown to influence the content sharing decision on social media (Chang et al., 2015).

Several studies have suggested that information-focused content may irritate customers in the context of online and social media marketing. Therefore, the sharing of such content could be limited (Tellis et al., 2019). One of the reasons quoted in the literature for the resistance in sharing information-focused content is the lack of altruistic motives in sharing. These motives may be limited when the content focuses on factual features like names, brands, locations, and organizations instead of emotional goals that receivers can achieve.

However, given the COVID-19 context, this research uses the language expectancy theory to hypothesize that information-focused content could motivate content sharing. The theory posits that language is a rule-based system in which communicators and readers develop expected norms concerning communication styles in the given situations (Burgoon and Miller, 2018). This research draws upon the language expectancy theory and argues that information-driven content is expected during a pandemic situation. The reason could be attributed to the fact that a crisis like COVID-19 leads to a huge information gap and thus demands up-to-date, factual, and unambiguous information (Lee and Yu, 2020). In the situation of a pandemic, an individual's motivation is mainly for a utilitarian purpose. Thus, the degree of information could be a dominant predictor of information sharing and resharing behavior. There are three measures of informativeness (names, locations, and organizations mentioned in the post); therefore, three hypotheses are used to address the second research question, does information-focused content play a role in the sharing of COVID-19 content on Twitter?

H2a-H2c: *Tweets with more mentions of -Names (a), Locations (b), and Organizations (c) will have more retweets in the context of COVID-19 content.*

3.3 *Role of a tweet's linguistic features in COVID-19 content sharing*

Social media's role as an information source has gained prominence, especially in the face of emergencies (Lin et al., 2016a; Martinez-Rojas et al., 2018). Microblogging websites like Twitter provide functions such as creating and replying to posts, mentioning users, and retweeting options. The flexibility in resharing or retweeting the content has made Twitter the most preferred medium for information dissemination. Twitter tweets have a word limit of 280 characters (Isaac, 2017) and can include URL links, user mentions, and hashtags. Therefore, the linguistic features of Tweets in this format could offer interesting insights into the retweeting behavior of users, especially in the context of COVID-19. Lexical features of a tweet, like its length and content features, such as the usage of hashtags, user mentions, and the number of followers, could play a significant role in deciding the speed and spread of information content.

Literature has demonstrated the influence of a tweet's content features such as hashtags, URL, user mentions, and contextual features like the number of followers on retweet probability (Suh et al., 2010). Zarrella (2009) extended the work by analyzing linguistic, psychological, and timing features along with URL, hashtag, and the number of followers of a user. A systematic attempt to assess the motivations behind the usage of hashtags was conducted by Rauschnabel et al. (2019). The authors uncovered ten different motivations for hashtag use, namely amusing, organizing, designing, conforming, trend gaging, bonding, inspiring, reaching, summarizing, and endorsing.

Since this research explores the factors affecting the retweet of COVID-19 tweets, it is important to understand the factors that affect individual content-sharing behavior in a crisis. Studies have explored major features that predict sharing and retweeting possibilities of a tweet during disasters and crises (Li et al., 2018; Lin et al., 2016a). The diffusion of information via Twitter during crises and the role of content features that ensure effective communication were also studied by Son et al. (2019). The study concluded that the average retweet time decreased with an increase in the number of words and decreased with the inclusion of a URL. Bruns and Stieglitz (2012) found that tweets trending during natural disasters used more URLs than user designations. Such tweets had a higher retweet count than those from the mainstream media event. Hashtags also play a vital role in retrieving essential information from the millions of tweets and retweets generated during disasters, according to L.Potts et al. (2011). This research argues that the length of a tweet could increase the probability of retweet count. It also posits that using many hashtags and user mentions in a tweet could be negatively associated with retweet count.

Retweeting or reposting behaviors of a user plays a crucial role in information dissemination via Twitter (Kwon et al., 2017). According to Lee and Yu (2020), one of the significant features contributing towards higher retweet count is word count or length of a tweet. During a crisis, information-rich content has a higher chance of being retweeted. The authors defined the word count of a tweet as an essential measure for evaluating language richness. Hence, it can be argued that an increase in the length of the tweet positively impacts the retweet count, as this helps in disseminating pandemic-related guidelines and medical information with clarity.

When evaluating hashtags on Twitter, their quality is an important aspect to be considered. Millions of hashtags are generated in response to a crisis or disaster. Too many of them can hinder a user's ability to process information. Therefore, we emphasize the need for careful selection of influential hashtags to ensure increased visibility of a tweet and chances of a retweet. Usernames are extensively used on Twitter search engines to search for a profile or tweets mentioning that particular user. The tweet's author enables more interactions by drawing further attention to their tweet by using @username (Tang et al.,2014). This can help increase the visibility of a profile or a brand. However, in the context of COVID-19, over usage of user mentions can negatively impact the retweet possibilities. In a pandemic event, users prefer information intended towards the public than to a specific user or a specific set of users. Hence, we argue that an increase in username mentions can reduce a tweet's visibility and retweet possibilities.

Therefore, based on the three factors (tweet length, number of hashtags, and user mentions), three hypotheses are used to address the third research question, Do linguistic features of a tweet play a role in the sharing of COVID-19 content on Twitter?

H3 a. The number of words in a tweet is positively associated with the retweet count

H3 b. An increase in the number of hashtags in a tweet is negatively associated with the retweet count

H3 c. An increase in the number of user mentions negatively impacts the retweet count.

4. Research Method

4.1 Data Overview

This research uses secondary data to test the conceptual framework. The tweets were collected from 29th March 2020 till 29th April 2020 (one month). The dataset is available in the public domain (Smith, 2020), and has been used by several studies in the context of COVID-19 (Kruspe et al., 2020; Zarei et al., 2020). The extractions were done using Twitter's streaming Application Programming Interface (API) (Twitter, 2020) with the help of the 'rtweet' package on CRAN. The following hashtags were used to filter the tweets to be included in the dataset: #coronavirus, #coronavirusoutbreak, #coronavirusPandemic, #covid19, #covid_19 and #ihavecorona. The data extraction was executed at the end of each day to maintain its consistency.

The Twitter data of each day was stored in a separate file as there were 500,000 plus tweets per day on average. All the sheets were combined onto a single dataset with the help of the big data platform on Microsoft Azure Machine Learning Studio. Since the research objective was to understand the factors that impact the sharing of content, retweet count was the critical variable in the dataset. It was decided to consider only the tweets that were tweeted in a common time window of the day (15 minutes), allowing a sufficient number of hours to reach the audience and examine the retweet. This step ensured fairness in the comparison of the tweets on retweet count. Several steps were executed to clean the dataset - treatment of missing values, examining inappropriate data points, and handling invalid attributes. The repetition of tweets in the dataset was avoided by removing retweets, and only the first occurrence of a tweet was retained (original tweet). The language filter was set to English as most of the machine learning packages do not apply to other languages. The entire filtering process resulted in a substantial reduction in the number of tweets recorded in the final dataset (57, 331 tweets).

4.2 *Research method overview and coding process*

The final dataset was subjected to the coding process to extract variables that could be used for further analysis. The text in the Tweets had to be processed for quality, and the Azure Machine Learning studio was used to make the tweets eligible for natural language processing. The stop words were removed, the case was normalized to lower case, duplicate characters were removed, URLs were removed, and the expansion of verb contraction was done. After preprocessing of text, the following coding process was put in place: emotion analysis (as part of sentiment), topic modeling, Named Entity Recognition (NER), and other linguistic features. The resulting dataset was subjected to linear regression to test the hypotheses and predict the retweet count. The process is summarized in Figure-1.

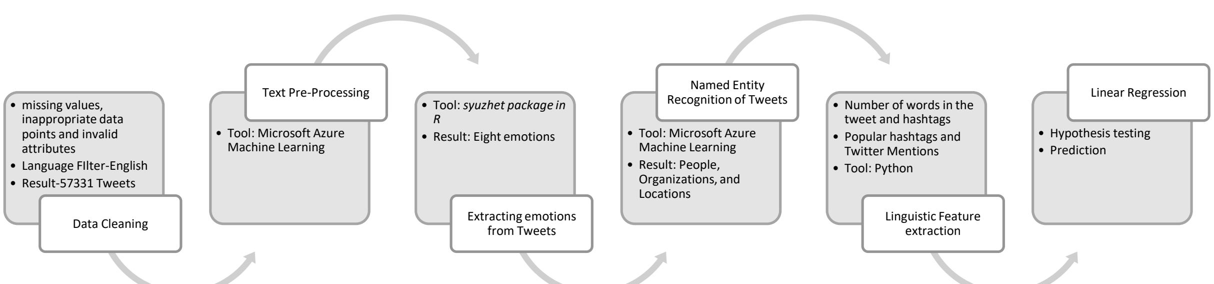

Figure 1: Research method overview and coding process

4.3 Measures and variables in the model

Retweet Count: The retweet count was captured in the dataset and served as the dependent variable for the analysis. The Shapiro–Wilk W test for normality revealed that the retweet count was not normally distributed ($p<.00001$). Therefore, the log transform of the retweet count was used for further analysis.

Coding emotions: Analyzing the emotions, mood, and sentiments from a text are usually categorized under the common umbrella of sentiment analysis (Alessia et al., 2015). It has been applied in various contexts in the literature, and the topics range from political sentiments (Murthy, 2015) to product review sentiment and emotions (Feldman, 2013). Two key outcomes of sentimental analysis could be useful for this study, polarity of sentiments (positive and negative) and the extent of emotions (trust, surprise, anger, and others) in a text (Taboada et al., 2011). This study used the extent of emotions to understand their influence on the sharing of content.

The analysis level to extract the text's emotions can be conducted at three levels document, sentence, and aspect level (Feldman, 2013). Each Tweet is considered a document in this research, and thus, the analysis is conducted at the document level. There are two approaches to sentiment analysis in terms of the methodology – lexicon-based and machine learning methods (Liu and Lei, 2018). The former method uses the sentiment lexicon containing a list of target words to meet the objective, while the latter uses training and test datasets to determine the semantic orientation of the text. Since the machine learning approach has limited applicability across diverse domains (Taboada et al., 2011), and there is a lack of readily available data, the lexicon-based method was used for this research.

There are several lexicon-based packages available to extract the emotions from the text. The authors chose R (Version 3.6.2) as a tool and *syuzhet* package by Jockers (2017) to execute the sentiment analysis. The corpora were submitted to a document-level (each tweet), lexicon-based analysis using the package in R. Several studies have used the *syuzhet* package with big data in a variety of domains (France and Ghose, 2019; Choudhury et al., 2019; Heinig and Nanda, 2018). The outcome of this package is in the form of eight emotions for each tweet- “anger,” “ fear,” “anticipation,” “trust,” “surprise,” “sadness,” “joy,” and “disgust.”. These emotions have been well-documented in the psychology literature (Plutchik, 1994). The emotional analysis of the tweets by the package was executed with the help of the National Research Council Canada (NRC) word-emotion lexicon (Mohammad and Turney, 2013). A summary of these emotions is provided in Table-2.

Table 2. Summary of emotions labeled on the Twitter dataset.

Emotion	**Count of occurrences (across 57331 tweets)**	**Percentage**
Anger	20718	9.97
Anticipation	30900	14.88
Disgust	12338	5.94
Fear	39817	19.17
Joy	18291	8.81
Sadness	26938	12.97
Surprise	13714	6.60
Trust	45013	21.67

Topic Modeling: Other than the factors included in the conceptual model, it was decided to test for specific topics that could influence the sharing of COVID-19 tweets. The tweets in the dataset were, therefore, subjected to topic modeling. Since the topics were completely unknown before the analysis, they could not be included in the conceptual model but were tested in the prediction model. The field of topic modeling includes several algorithms that can process vast volumes of text data to analyze and identify the dominant themes (Rathore et al., 2017).

There are several methods discussed in the literature for extracting topics from a short text. Yang et al. (2019) classify these methods into representation-based and model-based methods. Representation-based methods focus on learning effective vectors for text and exploiting clustering methods like K-Means (Jain, 2010). While artificial intelligence algorithms are used at one end, extracting the latent semantic for text documents is another popular method. The latter methods include Latent Semantic Index (LSI), which uses a term-document matrix (Deerwester et al., 1990), and Latent Dirichlet Allocation (LDA), which assumes that text documents are generated by a set of latent topics (Blei et al., 2003). On the other hand, model-based methods focus on new clustering strategies and do not need to represent short text with feature vectors (Li et al., 2016; Yin and Wang, 2014).

The Latent Dirichlet Allocation algorithm was selected in this study to execute topic modeling. This method can control the number of words and topics that could ease conducting the empirical analysis. The tool selected for this process was Microsoft Azure Machine Learning Studio due to its advanced ability to handle big data and pre-process text. As an output of topic modeling, the studio generates two outputs, the feature-topic matrix and topic loading for each tweet. The two outputs could be interpreted well if the correct number of topics (N) were chosen to receive meaningful results (Law and Jain, 2003). The number of training iterations was set to 1000 as the topic structure appeared more stable. The number of words in the topic was under ten to attain a balance between interpretation ability and sufficiency of words. The authors experimented with various values of N in the recommended range of 5 to 30 (Chen et al., 2016). The results were then traced back to the literature to understand the influence of sharing content on social media. Three domain experts were also consulted to examine the corpus and number of topics. The method has been used by studies that deploy topic modeling for conducting empirical analysis (Kar, 2020). A summary of model parameters used for the topic modeling is provided in Table 3a.

Table 3a. Model parameters

Parameter	Value
Latent Dirichlet Allocation	
Number of training iterations	1000
Rho parameter	0.01
Alpha parameter	0.01
Normalize	True
Number of hash bits	12
Emotions coding	
Dictionary	NRC Sentiment Dictionary

Based on the relevance of topics with the literature on content sharing and expert inputs, five topics were chosen to record each tweet's loadings across them. The choice of N-grams was fixed to two (bigrams) because tweets have limited characters and are relatively shorter. Feature–topic matrix resulted in more than 180,000 features that were loaded on five topics. The top loadings of each topic were examined by reading the features to cluster them by themes. A summary of five topics is presented in Table-3b. The final dataset was appended with five new variables (five topic loadings), and each value ranged from 0 to 1. A value closer to 1 indicated a high loading of a tweet on a particular topic.

Table 3b. Summary of Topic Modeling

Topic Number	Topic Label	Keywords
Topic 1	Home isolation/ Stay at home	Home coffee; conference call; miss them; #workfromhome; homeworkout; #stayathome; home loackdown; hiding truth; #shelterinplace
Topic 2	Hygiene and sanitization	Handsanitizer donate; sanitization; the disinfectant; wash hands; more cleaning; have clean.
Topic 3	Government and Policies	Local government; its #government; our governments; #PandemicPolicy #nation; national leader; social-distancing rules; protecting rules.
Topic 4	Mental health	#domesticabuse; more stressful; very distressing; or suicide; worry about; #PanicDepression; about depression.
Topic 5	Optimism and solution	#LetsDoThisTogether; #wecandothis; vaccine trial; #Covid_19 Hope; #brightspot #hope; #LetsFightCovid19; #weareinthistogether

Coding linguistic features of the Tweets: Text mining helps to extract hidden meaningful insights from text data that further accelerate knowledge discovery (Kobayashi et al.,2018). Here, the authors have extracted the linguistic features of a tweet, as they play a major role in influencing its virality. The major features extracted are the length (number of words), the number of hashtags present, and the number of users mentions in a tweet. The authors chose Python 3 environment to extract linguistic features, as it has several inbuilt analytics libraries such as nltk, used in this work. Python is an efficient tool that can facilitate the analysis and fine cleaning of data (Sun, 2020), which includes removing stop words, punctuations, and special characters that are generally present in tweets.

Coding information-focused content: To extract the content that displayed information-focused content and facts, Named Entity Recognition (NER) was used. It is a subproblem of information extraction and usually involves analyzing documents and identifying expressions that refer to people, places, organizations, and brands (Mansouri et al., 2008). This method is also popularly known as Entity Identification, Entity Extraction, or Entity Chunking. The method in the literature has been applied in various domains. It started with analyzing mentions from journalistic articles and then turned into analytics in other domains like biomedical text (Metke-Jimenez and Karimi, 2015) and queries (Guo et al., 2009). Several authors have used NER in the context of social media research, and thus, it can be considered as an effective method for tweet analysis (Aguilar et al., 2019; Phan and Sun, 2018).

The Named Entity Recognition was executed in Microsoft Azure Machine Learning studio as the module processes the text using both NER and Entity Linking capabilities. It, therefore, makes it a powerful solution to extract structured information from unstructured text (Makadia, 2020). The platform identifies three types of entities: people (PER), locations (LOC), and organizations (ORG). Therefore, for each tweet in the dataset, three entities (count) were recorded to be further used for analysis. A summary of all the variables is provided in Table-4.

Table 4. Summary of variables in the regression model

Variable	**Nature of the Variable**	**Comments**	**Summary Statistics**
Ln (Retweet_Count) (Dependent)	Log transform of the retweet count	All the tweets were selected from a common timeframe to ensure similar reach times.	
Anger	*Count of occurrence*. Numbers provided in Table-2.	Extracted from Syuzhet package in R.	Min: 0, Max: 7, Mean: 0.43
Anticipation			Min: 0, Max: 6, Mean: 0.65
Disgust			Min: 0, Max: 10, Mean: 0.26
Fear			Min: 0, Max: 8, Mean: 0.84
Joy			Min: 0, Max: 6, Mean: 0.38
Sadness			Min: 0, Max: 7, Mean: 0.59
Surprise			Min: 0, Max: 5, Mean: 0.28
Trust			Min: 0, Max: 9, Mean: 0.95
Topic 1	Topic Loading of each news article ranging from 0 to 1	LDA topic modeling using five topics and bi-grams. Tool: Azure Machine Learning	Max loading: 0.92
Topic 2			Max loading: 0.96
Topic 3			Max loading: 0.99
Topic 4			Max loading: 0.99
Topic 5			Max loading: 0.89
PER (People)	Count of names/people in the tweet	Extracted from Named Entity Recognition (NER) using Azure Machine Learning Studio	Min:0, Max: 8, Mean: 0.25
LOC (Location)	Count of locations in the tweet		Min:0, Max: 16 Mean: 0.45
ORG (Organization)	Count of organizations in the tweet		Min:0, Max: 10, Mean: 0.14
NWords_tweet	Number of words in a Tweet	Extracted from Tweets using Python	Min: 22,Max: 932 ,Mean: 228
N_Hashtags	Number of hashtags in a Tweet		Min: 0,Max: 29, Mean: 2.29,Mode: 1
N_Mentions	Number of Twitter Profile Mentions (@)		Min:0,Max:50,Mean: 0.62,Mode: 0
Follower_Count	Count of followers – Each user	Used as a control variable	Min:0, Max: 83741424, Mean: 81851.9

Data Analysis Strategies

The design of this research in order to test the conceptual model includes a mixed-effects linear regression model. The mixed-effects model has been popular in the context of social media content sharing studies and used in the contexts of ads and brands (Boksem and Smidts, 2015; Tellis et al., 2019). In this model, the log of retweet count was regressed on twenty independent variables listed in Table-4. The regression equation can be represented as:

$$\ln(Retweet\ Count) = \alpha_{Tweet} + \sum_{i=1}^{8} b_i Emotion_i + \sum_{i=1}^{5} c_i Topic_i + \sum_{i=1}^{3} d_i NER_i + \sum_{i=1}^{4} e_i\, Linguistic_i + Follower\_Count + \epsilon$$

In the equation above, $Emotion_i$ represents the eight emotions, $Topic_i$ represents the five topic loadings extracted from topic modeling, NER_i includes the number of people, locations, and organizations extracted via Named Entity Recognition, while $Linguistic_i$ represents four linguistic features of the tweets. The terms α, b_i, c_i, d_i, and e_i are coefficients to be estimated, and ϵ are error terms initially assumed to be independent and identically distributed. The impact of tweet parameters was controlled in two ways for estimating the retweet count. First, the Twitter follower count of the user posting the tweet was extracted to account for the observed heterogeneity in the profile popularity. Some profiles may have more followers than others affecting the visibility of the tweets. Second, a tweet-level random intercept (α_{Tweet}) was included to account for any additional unobserved heterogeneity in the tweet or the profile that may influence content sharing.

Before proceeding with the analysis, the assumptions of multicollinearity were checked (Hilbe, 2009). It was tested if there is any strong relationship between the independent variables. The indicator used to assess this assumption was the Variance Inflation Factor (VIF). All the VIF scores in the independent variables used for the regression were lower than 7.4, and hence the assumption was not violated.

5. Results

Table-5 reports the estimated effects of all the independent variables (summarized in Table-4) on the sharing of COVID-19 content (retweet count). These values are the estimated effects on the log of retweets from the mixed-effects regression model. The table is divided into the following categories: emotion-focused content, information-focused content, topic modeling loadings, and linguistic features. The four categories are based on the conceptual model proposed in the earlier sections.

To get the best fit model, multiple models were tested and benchmarked using the mixed-effects model suggested by Nakagawa and Schielzeth (2013). Since the overall model accounts for both within-individual and between-individual variance, the linear regression model's conventional R2 values might not be suitable for model evaluation. The conditional R^2 value is the recommended parameter, and the results indicated a score of 0.148. Marginal R^2 reflected the variance due to the fixed effects, and the model demonstrated a score of 0.108. The overall regression model is significant with F=38.43 and p<.0001.

Emotion-focused content: There were eight emotions included as the independent variables in the model: anger, anticipation, disgust, fear, joy, surprise, sadness, and trust. It was hypothesized that all eight emotions (mix of positive and negative) would significantly impact the sharing of COVID-19 content on Twitter. The results of five variables: anger (b=.026, p<.01), disgust (b=.024, p<.01), fear (b=.004, p<.05), sadness (b=.021, p<.01) and surprise (b=.029, p<.001), confirmed hypotheses H1a, H1c, H1d, H1f and H1g. All the five variables that turned out to be significant were negative emotions except surprise that could be categorized as both positive and negative. The positive emotions (anticipation, joy, and trust) did not turn out to be significant predictors of retweets. Therefore, three hypotheses – H1b, H1e, and H1h were not supported in the context of emotion-focused content.

Information-focused content: There were three independent variables included in the category of information-focused content- count of names (people), organizations, and locations in each tweet. It was hypothesized that all three variables would significantly impact the sharing of COVID-19 content on Twitter. The results of all three variables: people (b=0.062, p<0.001), organization (b=.033, p<.01) and location (b=.021, p<.001) confirmed hypotheses H2a, H2b and H2c. Therefore, the factors extracted with Named Entity Recognition significantly affected the retweet count on Twitter.

Topic Modeling: Five topics were modeled with the help of the Latent Dirichlet Allocation (LDA) algorithm. These topics were analyzed from the corpora of COVID-19 tweets, and the five categories obtained were: Home isolation/ Stay at home (Topic 1); Hygiene and sanitization (Topic 2); Government and Policies (Topic 3); Mental health (Topic 4); and Optimism and solution (Topic 5). Two topics had a significant and positive effect on the retweet count of COVID-19 tweets: Topic 4 (b=.132, p<.001) and Topic 5 (b=.376, p<.001). Therefore, tweets with text content related to mental health and optimism had higher chances of being shared on Twitter.

Linguistic Features: Three linguistic features were explored in the context of COVID-19 tweets: NWords_tweet, N_Hashtags, and N_Mentions. One variable had a positive and significant impact on the retweet count: Nwords_tweet (b=.001, p<.001). Therefore, COVID-19 tweets with more words had significantly higher chances of being retweeted on Twitter. Two variables had significant negative impact on the retweet count: N_Hashtags (b=-.035, p<.001) and N_Mentions (b=-.032, p<.001). The results indicate that the inclusion of more hashtags and user mentions on Twitter could reduce the chances of retweets in the context of COVID-19 content. All the four variables combined demonstrated the support of three hypotheses proposed in the conceptual model: H3a, H3b, and H3c.

Table 5. Estimated effects of COVID-19 Tweet characteristics on retweet count from the Mixed-Effects Model

Variable	Beta Coefficient	Standard Error	p-Value
Emotion-focused content			
Anger	0.026	0.009	0.003**
Anticipation	-0.027	0.006	0.500
Disgust	0.024	0.010	0.010 **
Fear	0.004	0.007	0.050*
Joy	0.015	0.009	0.074
Sadness	0.021	0.008	0.009**
Surprise	0.029	0.009	0.001**
Trust	0.000	0.006	0.998

Information-focused content			
PER (People)	0.062	0.008	0.000***
LOC (Location)	0.021	0.005	0.000***
ORG (Organization)	0.033	0.010	0.002**
Topic Modeling			
Topic 1	-0.015	0.104	0.887
Topic 2	0.065	0.137	0.635
Topic 3	0.052	0.114	0.521
Topic 4	0.132	0.040	0.001***
Topic 5	0.376	0.044	0.000***
Linguistic Features			
NWords_tweet	0.001	0.000	0.000***
N_Hashtags	-0.035	0.003	0.000***
N_Mentions	-0.032	0.003	0.000***
Control Variable			
Follower_Count	.001	0.000	0.070

Notes: * p<.05; ** p<.01; *** p<0.001.

6. Discussion

Information dissemination is an important topic in the context of social media that has garnered much research interest. This research brings in the perspective of information dissemination in an uncertain situation like the COVID-19 pandemic. The research model's dependent variable was the retweet count. Reposting behavior on Twitter was considered a crucial mechanism for information dissemination (Kwon et al., 2017). The results revealed interesting insights to advance research in the area of information sharing during a pandemic situation. Several theories prevalent in literature could be used to throw light on the way content is shared in the context of a crisis like the current COVID-19. While the uncertainty reduction theory (Berger and Bradac, 1982) posits reducing uncertainty in sharing information, the language expectancy theory (Burgoon and Miller, 2018) describes the expected norms concerning communication in crises. Our results are consistent with existing literature and also provide new perspectives on content sharing.

Overall, the hypotheses in all three categories proposed in the conceptual model were supported. While the first set of hypotheses (H1, emotions) received partial support, the other two sets (H2 and H3) received overall support. As part of the topic modeling exercise, two topics turned out to significantly impact the retweet of COVID-19 content. The two topics were mental health and optimism/solution. The results are discussed under five main categories: emotion-focused content, information-focused content, linguistic features, and topic modeling.

The results of *emotion-focused content* are discussed using two categories, positive and negative. Literature has emphasized the differences between positive and negative emotions (Keltner and Lerner, 2010) which are often used in viral content examination. The conceptual model of this research proposed the positive influence of both positive and negative emotions on the sharing of COVID-19 content. However, the results revealed that only negative emotions had a significant impact on the retweet count. While the significance of this impact was conceptually derived from the inside-out perspective, the insignificance of positive emotions could be explained with the outside-in perspective of COVID-19 suggested by Nabity-Grover et al. (2020). This perspective finds the topics that people regularly share on social media might now be ignored or frowned upon. This could particularly be applicable for positive content that displays happy and positive actions like travel, food, entertainment, and fun. Such content might not have relevance for social media users, given the pandemic context like COVID-19. The only emotion that could be considered both positive and negative was the surprise motion that also turned out to be a significant retweet count predictor. The authors did further analysis with a bag of words to figure out the orientation of surprise-driven content and found the dominance of negative emotions.

The results of the *information-focused content* revealed results that could extend the debate in the literature. As discussed in the conceptual model, several studies have looked at information-rich (factual) content as a roadblock to better reach and impressions (Akpinar and Berger, 2017). Because the content involves attributes like names, locations, and organizations, the information posted could appear dull putting off users from sharing it. However, one of the arguments used in such a proposition is that information-focused content could target names and entities that are already well-known by the users. On the other hand, in the context of COVID-19, the information coming every day could carry new information even if there were facts and information-loaded content.

It is also interesting to observe the hypothesis of information-loaded content negatively influencing content sharing proposed by Tellis et al. (2019). This hypothesis, however, had an exception to high-risk conditions. COVID-19 can be considered a risky and uncertain situation (Brownstone et al., 2020), and hence the results of this study can be justified. The mixed-effects regression model results revealed that information-focused content (count of people, locations, and organizations) positively influences the retweet count of COVID-19 content. With the change in COVID guidelines and the economic uncertainty, the risk of consuming information is high, and the consequences are unknown. Users often search for information that reduces uncertainty when the risk is high (Locander and Hermann, 1979). In situations like these, information-focused content can provide facts and arguments to reduce risk perceptions. Similar thoughts have been shared in the marketing domain, wherein information-focused ads could generate more impact when the product is new, or the market is uncertain (Chandy et al., 2001; MacInnis et al., 2002). Some marketing studies that have found the opposite impact have not considered uncertainty and new conditions as moderators. Therefore, this research extends the importance of new and uncertain conditions (COVID-19) in the context of information loaded (NER) content sharing.

The *topic modeling exercise* revealed two significant COVID-19 topics that had a high probability of being shared – Mental health and Optimism. The first topic of mental health included issues like suicide, depression, domestic violence, and others. This topic has been very prominent in the context of COVID-19, and the pandemic has been associated with psychological distress and symptoms of mental illness (Hao et al., 2020). A review of articles conducted by Rajkumar (2020) on mental health and COVID-19 demonstrated the impact of this topic and how several researchers are discussing this issue. The issue of mental health and how people are posting about it on Twitter has also been discussed in the literature (Lwin et al., 2020). However, the crucial additional perspective offered by the results in this study is that people are not only posting about mental health issues but also sharing these posts for a wider reach to demonstrate support.

The second topic that had a significant impact on the retweet count was the act of demonstrating optimism in the face of COVID-19. It included issues and, more importantly, hashtags like #Weareinthistogether and #wecandothis. While few studies find using such hashtags as a catalyzer for relationship marketing opportunities in COVID-19 (Steinhoff and Palmatier, 2020), others see it as an act of optimism and action (Storz and Wynfield, 2020). The results indicate that people not only see hope and are optimistic about the situation but also tend to share such content more often than others.

It was interesting to observe that while the second topic dealt with positive emotion (optimism and solution) and turned out to be significant, positive emotion-focused content (anticipation, joy, and trust) did not turn out to be significant. We investigated the types of emotions discussed in the literature to understand the results. Several studies have used an arousal perspective to classify emotions into high arousal and low arousal categories (Berger and Milkman, 2012; Lim, 2016; Tsai, 2007). Based on the total strength with which primary emotions were felt, high arousal positive emotions included awe, anticipation, joy, and trust (Dobele et al., 2007). On the other hand, low arousal positive emotions included emotions like calmness, relaxation, optimism, and others. Therefore, the emotions included in the emotion-focused content that was not seen to be significant were the high-arousal emotions, while those of topic modeling (optimism and solution-oriented) were low arousal emotions which were significant predictors of the retweet count (pandemic context). This is a new perspective offered by our research in the context of COVID-19, where two types of emotions (low and high arousal) could result in different predictions of content virality. It also opens up new dimensions in the emotion-based content examination for virality as a future research direction.

The results from Linguistic features in this study revealed that tweets, including information-rich content, had a high probability of being retweeted. Locander and Hermann (1979) discussed how users prefer information-rich content during high-risk situations. A similar study by Lee and Yu (2020) also emphasizes how users prefer detailed and accurate information during a crisis. During high emergencies such as a pandemic, people often rely on social media to acquire information (González-Padilla et al.,2020) on health and safety protocols from government entities and other concerned authorities. Hashtags are one of the most widely used and useful Twitter features. Excessive use of hashtags and user mentions in tweets was revealed to influence the retweet count negatively. User mentions are used where tweets can exchange ideas or converse with specific individuals or groups (Suh et al., 2010). Hence, users tend to view tweets with user mentions as information aimed at specific individuals rather than the general public, thus reducing its reach and visibility. Similarly, excessive use of hashtags often makes information processing difficult for readers. Digital marketing thrives heavily on hashtags campaigns (Naraine et al., 2019) for better reach and visibility on online platforms. However, during a pandemic, unlike a normal scenario, users are more oriented towards obtaining relevant updates on the ongoing pandemic. With the amount of uncertainty involved and the rate at which events spiral out of control during a pandemic, the general public leans more towards tweets containing authentic information rather than those with several hashtags. Hence, excessive usage of hashtags can work against this cause.

7. Conclusion

In summary, this research explores a person's content-sharing behavior in the context of a pandemic like COVID-19. This paper built its theoretical foundation using linguistic features and theories such as the viral content theory, language expectancy theory, and uncertainty reduction theory. It was observed that negative emotions, linguistic features (hashtags, mentions, and others), mental health-related and optimistic content, and information-focused content impacted retweets made by users of the microblogging platform, Twitter. The research also demonstrates the changes in the content-sharing behavior of individuals in a crisis like COVID-19. The trends visible pre-pandemic are not the same during the pandemic, and new insights are revealed.

The study has both theoretical and practical implementations. As for practical applications, our results could give recommendations for officials and individuals who wish to use social media platforms to disseminate COVID-19-related information. The research could also contribute to practitioners. Since the model uses a mix of topic modeling, emotions, and linguistic features, a comprehensive approach to virality is considered. The approach is different from previous works that have considered these methods in isolation and offered rare insights on content-sharing in a pandemic situation like COVID-19. The research findings could encourage practitioners, volunteers, and organizations to focus on linguistic style cues and emotions for effective communication on social media. The information-focused content has a significant impact on the retweet count. This offers a clear perspective to the content creators in using the uncertainty reduction principle and including more facts to achieve a more extensive reach of their posts. It also suggests negative emotions could play an important role in content sharing as users understand these emotions (pain, grief, and disgust) during COVID-19 and are willing to share them for a wider reach and better social support.

The theoretical contribution of this study is two-fold. The article presents a multifaceted nature of information resharing behavior using a multi-theoretical approach. Previous studies have used thematic content analysis and manual coding procedures. Given the scale of text that needs to be analyzed, this research uses machine learning algorithms to automate the coding. This approach of natural language processing allows research in similar areas to overcome the disadvantages of manual coding. Second, drawing on essential theories related to content sharing, this research incorporates the significance of the crisis's timings and focuses on communication when a pandemic occurs. The results reveal that the sharing behavior of people changes in a crisis. Content that was widely shared earlier could be frowned upon now, and posts considered taboo could be shared more post-pandemic.

There are implications on understanding the overall aspect of Twitter users interacting during a pandemic. A classic anthropological interpretation of the increase in focus on observing the reactions of COVID-19 on social media would be to see tweeting as a practice of imitation. Literature has considered tweeting (under crisis) to have an observable social effect, where tweeting about an issue could affect that issue in practice (Breslin et al., 2020; Taussig 1992). Our research considers theoretical frameworks that focus not just on emotional and lexical aspects of the tweets but also on how these aspects change in a pandemic context. The language used in a tweet has an amplified effect on the world as they interact with technology and media (Barad, 2007). The effect is due to the characteristics of the language used in tweets, resulting in replicability, scalability (in the form of retweets), and searchability (Boyd, 2010). We explored the factors that could impact the retweets and thus investigated the change in theoretical underpinnings pre- and post-COVID-19.

In the broader context of Twitter users, we explored *what* topics people tweet in a pandemic context (topic modeling) and how they tweet (emotions, sentiments, and facts). The text analytics techniques used for modeling have been well accepted in the literature (Mohammad and Kiritchenko, 2015). The study of tweeting behavior would help understand the societal roles that differ in the way they exert influence on a Twitter communication network (Mirbabaie et al., 2020). Various individual and organizational entities affect public discussion on Twitter through authoring and retweeting. The study of visibility and sharing of these discussions could influence sense-making. The notion could be relevant to crisis communication and organizational social media usage and open up new dimensions of sense-making and influence. One of the prominent theoretical concepts that got validated in our study was the inside-out and outside-in perspective (Nabity-Grover et al., 2020). It explains the shifting perception of what constitutes sensitive and private information. This could have significant implications on data protection practices, privacy rules, and establishing content-sharing norms on social media.

There are a few limitations to this study. First, the pandemic situation is discussed in the context of COVID-19 and does not include any other crisis events and situations. Therefore, generalizing the results to a broader context of risky and critical conditions based on linguistics could be difficult. Second, the study included only tweets from one social media platform (Twitter). Other platforms have different rules for posting content, and hence the behavior of sharing could differ. Future studies could extend the current research by using multiple social media platforms and incorporating various pandemic stages. Analyzing various locations could also offer geography-specific content sharing behavior.

References

Abbas, M. A., Eliyana, A., Ekowati, D., Saud, M. M., Raza, M. A., & Wardani, M. R. (2020). Data set on coping strategies in the digital age: The role of psychological well-being and social capital among university students in Java Timor, Surabaya, Indonesia. *Data in Brief*, 105583.

Adamic, L. A., Huberman, B. A., Barabási, A. L., Albert, R., Jeong, H., & Bianconi, G. (2000). Power-law distribution of the world wide web. *science*, *287*(5461), 2115-2115.

Aguilar, G., Maharjan, S., López-Monroy, A. P., & Solorio, T. (2019). A multi-task approach for named entity recognition in social media data. *arXiv preprint arXiv:1906.04135*.

Ahmed, W., Vidal-Alaball, J., Downing, J., & Seguí, F. L. (2020). COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. *Journal of Medical Internet Research*, *22*(5), e19458.

Akpinar, E., & Berger, J. (2017). Valuable virality. *Journal of Marketing Research*, *54*(2), 318-330.

Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, *125*(3).

Allington, D., Duffy, B., Wessely, S., Dhavan, N., & Rubin, J. (2020). Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. Psychological medicine, 1-7.

Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology* (Vol. 1, pp. 492-499). IEEE.

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. duke university Press.

Berger, C. R., & Bradac, J. J. (1982). *Language and social knowledge: Uncertainty in interpersonal relations* (Vol. 2). Hodder Education.

Berger, J. (2011). Arousal increases social transmission of information. *Psychological science*, *22*(7), 891-893.

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

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Boksem, M. A., & Smidts, A. (2015). Brain responses to movie trailers predict individual preferences for movies and their population-wide commercial success. *Journal of Marketing Research*, *52*(4), 482-492.

Boyd, D. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. *Networked self: Identity, community, and culture on social network sites*, *39*, 58.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *2010 43rd Hawaii international conference on system sciences* (pp. 1-10). IEEE.

Breslin, S. D., Enggaard, T. R., Blok, A., Gårdhus, T., & Pedersen, M. A. (2020). How We Tweet About Coronavirus, and Why: A Computational Anthropological Mapping of Political Attention on Danish Twitter during the COVID-19 Pandemic. In *The COVID-19 Forum III. University of St. Andrews*.

Bright, J., Hale, S., Ganesh, B., Bulovsky, A., Margetts, H., & Howard, P. (2020). Does Campaigning on Social Media Make a Difference? Evidence from candidate use of Twitter during the 2015 and 2017 UK Elections. Communication Research, 47(7), 988-1009.

Brooks, A. W., & Schweitzer, M. E. (2011). Can Nervous Nelly negotiate? How anxiety causes negotiators to make low first offers, exit early, and earn less profit. *Organizational Behavior and Human Decision Processes*, *115*(1), 43-54.

Brown, J. (2020). Social media shaming is spiking during the coronavirus pandemic, for better or worse. The Colorado Sun April 1. https://coloradosun.com/2020/04/01/socialmedia-shaming-about-coronavirus/

Brownstone, N. D., Thibodeaux, Q. G., Reddy, V. D., Myers, B. A., Chan, S. Y., Bhutani, T., & Liao, W. (2020). Novel coronavirus disease (COVID-19) and biologic therapy in psoriasis: infection risk and patient counseling in uncertain times. *Dermatology and therapy*, 1-11.

Bruns, A., & Stieglitz, S. (2012). *Quantitative Approaches to Comparing Communication Patterns on Twitter. Journal of Technology in Human Services, 30(3-4), 160–185.* doi:10.1080/15228835.2012.744249

Bruns, A., Burgess, J. E., Crawford, K., & Shaw, F. (2012). # qldfloods and@ QPSMedia: Crisis communication on Twitter in the 2011 south east Queensland floods.

Burgoon, M., & Miller, G. R. (2018). An expectancy interpretation of language and persuasion. In *Recent advances in language, communication, and social psychology* (pp. 199-229). Routledge.

Cashmore, P. (2009). YouTube: Why do we watch?. *CNN Tech*.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in Twitter: The million follower fallacy. *Icwsm*, *10*(10-17), 30.

Chandy, R. K., Tellis, G. J., MacInnis, D. J., & Thaivanich, P. (2001). What to say when: Advertising appeals in evolving markets. *Journal of marketing Research*, *38*(4), 399-414.

Chang, Y. T., Yu, H., & Lu, H. P. (2015). Persuasive messages, popularity cohesion, and message diffusion in social media marketing. *Journal of Business Research*, *68*(4), 777-782.

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016, April). Topic modeling for evaluating students' reflective writing: A case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 1-5).

Choudhury, P., Wang, D., Carlson, N. A., & Khanna, T. (2019). Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal*, *40*(11), 1705-1732.

Dafonte-Gómez, A. (2015). The key elements of viral advertising. from motivation to emotion in the most shared videos. *arXiv preprint arXiv:1505.02002*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information systems research*, *3*(1), 60-95.

Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A., & Larson, H. (2020). The pandemic of social media panic travels faster than the COVID-19 outbreak.

Dobele, A., Lindgreen, A., Beverland, M., Vanhamme, J., & Van Wijk, R. (2007). Why pass on viral messages? Because they connect emotionally. *Business Horizons*, *50*(4), 291-304.

Edell, J. A., & Burke, M. C. (1987). The power of feelings in understanding advertising effects. *Journal of Consumer research*, 14(3), 421-433.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1998). Gift exchange and reciprocity in competitive experimental markets. *European Economic Review*, *42*(1), 1-34.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89.

Ferree, M. M., Gamson, W. A., Rucht, D., & Gerhards, J. (2002). *Shaping abortion discourse: Democracy and the public sphere in Germany and the United States*. Cambridge University Press.

Fox, L. (2020). COVID fears spawn social media shaming. The Alpena News April 22. https://www.thealpenanews.com/news/local-news/2020/04/covid-fears-spawnsocial-media-shaming/, (Accessed 5 June 2020).

France, S. L., & Ghose, S. (2019). Marketing analytics: Methods, practice, implementation, and links to other fields. *Expert Systems with Applications*, *119*, 456-475.

Gaertner, S. L., & Dovidio, J. F. (1977). The subtlety of White racism, arousal, and helping behavior. *Journal of Personality and social Psychology*, *35*(10), 691.

Gamson, W. A., & Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology*, *95*(1), 1-37.

Gao, J., Zheng, P., Jia, Y., Chen, H., Mao, Y., Chen, S., ... & Dai, J. (2020). Mental health problems and social media exposure during COVID-19 outbreak. Plos one, 15(4), e0231924.

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., ... & Verlegh, P. (2005). The firm's management of social interactions. *Marketing letters*, *16*(3-4), 415-428.

González-Padilla, Daniel A., & Tortolero-Blanco, Leonardo. (2020). Social media influence in the COVID-19 Pandemic. International braz j urol, 46(Suppl. 1), 120-124. Epub July 27, 2020.https://doi.org/10.1590/s1677-5538.ibju.2020.s121

Guo, J., Xu, G., Cheng, X., & Li, H. (2009, July). Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 267-274).

Hagerstrom, A., Alhabash, S., & Kononova, A. (2014). Emotional dimensionality and online ad virality: Investigating the effects of affective valence and content arousingness on processing and effectiveness of viral ads. In *American Academy of Advertising. Conference. Proceedings (Online)* (p. 109). American Academy of Advertising.

Han, Y., Lappas, T., & Sabnis, G. (2020). The Importance of Interactions Between Content Characteristics and Creator Characteristics for Studying Virality in Social Media. Information Systems Research.

Hao, X., Zhou, D., Li, Z., Zeng, G., Hao, N., Li, E., ... & Yan, B. (2020). Severe psychological distress among patients with epilepsy during the COVID-19 outbreak in southwest China. *Epilepsia*.

Harris, M. (2020). Coronavirus has amplified online shaming — Not just for influencers. Insider April 28. https://www.insider.com/coronavirus-shaming-viral-callout-onlinepandemic-behavior-covid-2020-4, (Accessed 5 June 2020).

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

Heinig, S., & Nanda, A. (2018). Measuring sentiment in real estate–a comparison study. *Journal of Property Investment & Finance*.

Hilbe, J. M. (2009). *Logistic regression models*. CRC press.

Hopp, T., Ferrucci, P., & Vargo, C. J. (2020). Why Do People Share Ideologically Extreme, False, and Misleading Content on Social Media? A Self-Report and Trace Data–Based Analysis of Countermedia Content Dissemination on Facebook and Twitter. *Human Communication Research*.

Hsieh, J. K., Hsieh, Y. C., & Tang, Y. C. (2012). Exploring the disseminating behaviors of eWOM marketing: persuasion in online video. *Electronic Commerce Research*, *12*(2), 201-224.

Hughes, A. L., St. Denis, L. A., Palen, L., & Anderson, K. M. (2014). Online public communications by police & fire services during the 2012 Hurricane Sandy. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1505-1514).

Hwang, Y., & Shim, H. (2010). Opinion leadership on Twitter and Twitter use. *Journal of Korean Broadcasting*, *24*(6), 365-404.

Isaac, M. (2017). Twitter to test doubling tweet length to 280 characters. *The New York Times*, *26*.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

Jockers, M. (2017). Package 'syuzhet'. *URL: https://cran. r-project. org/web/packages/syuzhet*.

Kar, A. K. (2020). What Affects Usage Satisfaction in Mobile Payments? Modelling User Generated Content to Develop the "Digital Service Usage Satisfaction Model". *Information Systems Frontiers*, 1-21.

Keltner, D., & Lerner, J. S. (2010). Emotion. To appear in *The Handbook of Social Psychology(5th edition) (D. Gilbert, S. Fiske, and G. Lindsey, Eds.)*.New York: McGraw Hill.

Kim, L., Fast, S. M., & Markuzon, N. (2019). Incorporating media data into a model of infectious disease transmission. PloS one, 14(2), e0197646.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. Organizational research methods, 21(3), 733-765.

Kordzadeh, N., & Warren, J. (2017). Communicating personal health information in virtual health communities: an integration of privacy calculus model and affective commitment. *Journal of the Association for Information Systems*, 18(1), 1.

Kruspe, A., Kersten, J., & Klan, F. (2020). Detection of informative tweets in crisis events. *Natural Hazards and Earth System Sciences Discussions*, 1-18.

Kwon, J., Han, I., & Kim, B. (2017). Effects of source influence and peer referrals on information diffusion in Twitter. *Industrial Management & Data Systems*.

L. Potts, J. Seitzinger, D. Jones, and A. Harrison, "Tweeting Disaster: Hashtag Constructions and Collisions," Proc. 29th ACM International Conference on Design of communication, pp. 235–240, 2011

Larson, H. J. (2018). The biggest pandemic risk? Viral misinformation. Nature, 562(7726), 309-310.

Law, M. H., & Jain, A. K. (2003). Cluster validity by bootstrapping partitions. *Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, Tech. Rep. MSUCSE-03-5*.

Lee, C.H. and Yu, H. (2020), "The impact of language on retweeting during acute natural disasters: uncertainty reduction and language expectancy perspectives", *Industrial Management & Data Systems*, Vol. 120 No. 8, pp. 1501-1519.

Lee, E. J., Lee, H. Y., & Choi, S. (2020). Is the message the medium? How politicians' Twitter blunders affect perceived authenticity of Twitter communication. Computers in Human Behavior, 104, 106188.

Lee, J., & Hong, I. B. (2016). Predicting positive user responses to social media advertising: The roles of emotional appeal, informativeness, and creativity. *International Journal of Information Management*, *36*(3), 360-373.

Lee, M., Kim, H. Kim.(2012). Psychological motivations to retweeting: Attitudes, subjective norms and communicative motivations. *Korean Journal of Broadcasting*, *26*(4), 137-172.

Lee, M., Kim, H., & Kim, O. (2015). Why do people retweet a Tweet?: altruistic, egoistic, and reciprocity motivations for retweeting. *Psychologia*, *58*(4), 189-201.

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016, July). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 165-174).

Li, L., Zhang, Q., Tian, J. and Wang, H. (2018), "Characterizing information propagation patterns in emergencies: a case study with Yiliang Earthquake", International Journal of Information Management, Vol. 38 No. 1, pp. 34-41.

Lim, N., 2016. Cultural differences in emotion: differences in emotional arousal level between the East and the West. Integrative medicine research, 5(2), 105-109.

Lin, W. Y., Zhang, X., Song, H., & Omori, K. (2016a). Health information seeking in the Web 2.0 age: Trust in social media, uncertainty reduction, and self-disclosure. *Computers in Human Behavior*, *56*, 289-294.

Lin, X., Lachlan, K.A. and Spence, P.R. (2016b), "Exploring extreme events on social media: a comparison of user reposting/retweeting behaviors on Twitter and Weibo", Computers in Human Behavior, Vol. 65, pp. 576-581.

Liu, D., & Lei, L. (2018). The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election. *Discourse, context & media*, *25*, 143-152.

Locander, W. B., & Hermann, P. W. (1979). The effect of self-confidence and anxiety on information seeking in consumer risk reduction.

Lwin, M. O., Lu, J., Sheldenkar, A., Schulz, P. J., Shin, W., Gupta, R., & Yang, Y. (2020). Global sentiments surrounding the COVID-19 pandemic on Twitter: analysis of Twitter trends. *JMIR public health and surveillance*, *6*(2), e19447.

MacInnis, D. J., Rao, A. G., & Weiss, A. M. (2002). Assessing when increased media weight of real-world advertisements helps sales. *Journal of Marketing Research*, *39*(4), 391-407.

Maheswaran, D., & Meyers-Levy, J. (1990). The influence of message framing and issue involvement. *Journal of Marketing research*, *27*(3), 361-367.

Makadia, A. (2020). Public preview: Named Entity Recognition in the Cognitive Services Text Analytics API. [online]. Available https://azure.microsoft.com/en-us/blog/public-preview-named-entity-recognition-in-the-cognitive-services-text-analytics-api/

Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition approaches. *International Journal of Computer Science and Network Security*, *8*(2), 339-344.

Martinez-Rojas, M., del Carmen Pardo-Ferreira, M. and Rubio-Romero, J.C. (2018), "Twitter as a tool for the management and analysis of emergency situations: a systematic literature review", International Journal of Information Management, Vol. 43, pp. 196-208.

Mayrhofer, M., Matthes, J., Einwiller, S., & Naderer, B. (2020). User generated content presenting brands on social media increases young adults' purchase intention. *International Journal of Advertising*, *39*(1), 166-186.

McCauley, M., Minsky, S., & Viswanath, K. (2013). The H1N1 pandemic: media frames, stigmatization, and coping. BMC Public Health, 13(1), 1116.

McDonnell, T. E., Bail, C. A., & Tavory, I. (2017). A theory of resonance. *Sociological Theory*, *35*(1), 1-14.

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis: Can we trust what we RT? In *Proceedings of the first workshop on social media analytics* (pp. 71-79).

Merchant, R. M., & Lurie, N. (2020). Social media and emergency preparedness in response to novel coronavirus. *Jama*.

Metke-Jimenez, A., & Karimi, S. (2015). Concept extraction to identify adverse drug reactions in medical forums: A comparison of algorithms. *arXiv preprint arXiv:1504.06936*.

Mirbabaie, M., Bunker, D., Stieglitz, S., Marx, J., & Ehnis, C. (2020). Social media in times of crisis: Learning from Hurricane Harvey for the coronavirus disease 2019 pandemic response. *Journal of Information Technology*, *35*(3), 195-213.

Mohammad, S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, *31*(2), 301-326.

Mohammad, S. M., & Turney, P. D. (2013). Nrc emotion lexicon. *National Research Council, Canada*, *2*.

Murthy, D. (2015). Twitter and elections: are tweets, predictive, reactive, or a form of buzz? *Information, Communication & Society*, *18*(7), 816-831.

Nabity-Grover, T., Cheung, C. M., & Thatcher, J. B. (2020). Inside out and outside in: How the COVID-19 pandemic affects self-disclosure on social media. *International Journal of Information Management*, 102188.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, *4*(2), 133-142.

Naraine, M. L., Pegoraro, A., & Wear, H. (2019). # WeTheNorth: Examining an Online Brand Community Through a Professional Sport Organization's Hashtag Marketing Campaign. Communication & Sport, 2167479519878676.

Nelson-Field, K., Riebe, E., & Newstead, K. (2013). The emotions that drive viral video. *Australasian Marketing Journal (AMJ)*, *21*(4), 205-211.

Ni, M. Y., Yang, L., Leung, C. M., Li, N., Yao, X. I., Wang, Y., ... & Liao, Q. (2020). Mental health, risk factors, and social media use during the COVID-19 epidemic and cordon sanitaire among the community and health professionals in Wuhan, China: Cross-sectional survey. *JMIR mental health*, *7*(5), e19009.

Nikolinakou, A., & King, K. W. (2018). Viral video ads: Emotional triggers and social media virality. *Psychology & Marketing*, *35*(10), 715-726.

Oliveira, T., Araujo, B., & Tam, C. (2020). Why do people share their travel experiences on social media? *Tourism Management*, *78*, 104041.

Park, H. W., Park, S., & Chong, M. (2020). Conversations and medical news frames on Twitter: Infodemiological study on covid-19 in south korea. *Journal of Medical Internet Research*, *22*(5), e18897.

Park, N., & Jeong, J. (2011). Who retweet and why? Retweeting motivation factors and socio-demographic variables' influence on retweeting. *Journal of Media Economics and Culture*, *9*(3), 95-132.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychological science, 31(7), 770-780.

Peters, K., & Kashima, Y. (2007). From social talk to social action: shaping the social triad with emotion sharing. *Journal of personality and social psychology*, *93*(5), 780.

Pfefferbaum, B., & North, C. S. (2020). Mental health and the Covid-19 pandemic. *New England Journal of Medicine*.

Phan, M. C., & Sun, A. (2018, June). Conerel: Collective information extraction in news articles. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1273-1276).

Phelps, J. E., Lewis, R., Mobilio, L., Perry, D., & Raman, N. (2004). Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. *Journal of advertising research*, *44*(4), 333-348.

Plutchik, R. (1994). *The psychology and biology of emotion*. HarperCollins College Publishers.

Rajkumar, R. P. (2020). COVID-19 and mental health: A review of the existing literature. *Asian journal of psychiatry*, 102066.

Rathore, A. K., Kar, A. K., & Ilavarasan, P. V. (2017). Social media analytics: Literature review and directions for future research. *Decision Analysis*, *14*(4), 229-249.

Rauschnabel, P. A., Sheldon, P., & Herzfeldt, E. (2019). *What motivates users to hashtag on social media? Psychology & Marketing.* doi:10.1002/mar.21191

Rosenberg, H., Syed, S., & Rezaie, S. (2020). The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *Canadian Journal of Emergency Medicine*, 1-4.

Roy, K. C., Hasan, S., Sadri, A. M., & Cebrian, M. (2020). Understanding the efficiency of social media-based crisis communication during hurricane Sandy. International Journal of Information Management, 102060.

Rufai, S. R., & Bunce, C. (2020). World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. *Journal of Public Health*.

Savage, N. (2011). Twitter as medium and message. *Communications of the ACM*, *54*(3), 18-20.

Schudson, M. (1989). Perspectives from media studies on the efficacy of symbols. *Theory and Society*, *18*(2), 153-180.

Shirky, C. (2011). The political power of social media: Technology, the public sphere, and political change. *Foreign affairs*, 28-41.

Skaza, J., & Blais, B. (2017). Modeling the infectiousness of Twitter hashtags. Physica A: Statistical Mechanics and its Applications, 465, 289-296.

Smith (2020), "coronavirus (covid19) tweets", mar 2020. [online]. available: www.kaggle.com/smid80/coronavirus-covid19-tweets

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, *48*(4), 813.

Son, J., Lee, H.K., Jin, S. and Lee, J. (2019), "Content features of tweets for effective communication during disasters: a media synchronicity theory perspective", International Journal of Information Management, Vol. 45, pp. 56-68

St. Michel, P. (2020). Understanding the need to shame someone on social media for not exercising self-restraint during a pandemic. The Japan Times May 16. https://www. japantimes.co.jp/news/2020/05/16/national/media-national/social-mediashaming-coronavirus/, (Accessed 5 June 2020)

Stayman, D. M., & Batra, R. (1991). Encoding and retrieval of ad affect in memory. *Journal of Marketing Research*, 28(2), 232-23.

Steinhoff, L., & Palmatier, R. W. (2020). Commentary: Opportunities and challenges of technology in relationship marketing. *Australasian Marketing Journal (AMJ)*.

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

Storz, L., & Wynfield, A. (2020). How to Sense a Pandemic: Curves, Models, and the Affective Allure of Flattening. *American Ethnologist website*.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010, August). Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network. In *2010 IEEE Second International Conference on Social Computing* (pp. 177-184). IEEE.

Sun, G. (2020). Symmetry Analysis in Analyzing Cognitive and Emotional Attitudes for Tourism Consumers by Applying Artificial Intelligence Python Technology. *Symmetry*, *12*(4), 606.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

Tang, L., Ni, Z., Xiong, H., & Zhu, H. (2014). *Locating targets through mention in Twitter. World Wide Web, 18(4), 1019–1049.* doi:10.1007/s11280-014-0299-8.

Taussig, M. 1992. *Mimesis and Alterity: A Particular History of the Senses*. London: Routledge.

Teixeira, T., Wedel, M., & Pieters, R. (2012). Emotion-induced engagement in internet video advertisements. *Journal of marketing research*, *49*(2), 144-159.

Tellis, G. J., MacInnis, D. J., & Tirunillai, S. (2020). What Drives Virality (Sharing, Spread) of YouTube Video Ads: Emotion vs Brand Prominence and Information. Marketing Science Working Paper Series.

Tellis, G. J., MacInnis, D. J., Tirunillai, S., & Zhang, Y. (2019). What drives virality (sharing) of online digital content? The critical role of information, emotion, and brand prominence. *Journal of Marketing*, *83*(4), 1-20.

Tsai, J.L., 2007. Ideal affect: Cultural causes and behavioral consequences. Perspectives on Psychological Science, 2(3), pp.242-259.

Twitter. (2020). Consuming streaming data URL: https://developer.Twitter.com/en/docs/tutorials/consuming-streaming-data [accessed 2020-04-10]

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86-95.

Wang, B., & Zhuang, J. (2017). Crisis information distribution on Twitter: a content analysis of tweets during Hurricane Sandy. *Natural hazards*, *89*(1), 161-181.

Wang, Y., & Yang, Y. (2020). Dialogic communication on social media: How organizations use Twitter to build dialogic relationships with their publics. Computers in Human Behavior, 104, 106183.

Wilson, M. E., & Chen, L. H. (2020). Travellers give wings to novel coronavirus (2019-nCoV).

Wojnicki, A. C., & Godes, D. (2008). Word-of-mouth as self-enhancement. *HBS marketing research paper*, (06-01).

Xu, W. W., & Zhang, C. (2018). Sentiment, richness, authority, and relevance model of information sharing during social Crises—the case of# MH370 tweets. *Computers in Human Behavior*, *89*, 199-206.

Yang, S., Huang, G., & Cai, B. (2019). Discovering topic representative terms for short text clustering. *IEEE Access*, *7*, 92037-92047.

Yang, S., Jiang, J., Pal, A., Yu, K., Chen, F., & Yu, S. (2020). Analysis and Insights for Myths Circulating on Twitter During the COVID-19 Pandemic. *IEEE Open Journal of the Computer Society*, *1*, 209-219.

Yin, J., & Wang, J. (2014, August). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 233-242).

Yin, H., Yang, S., & Li, J. (2020, November). Detecting topic and sentiment dynamics due to Covid-19 pandemic using social media. In *International Conference on Advanced Data Mining and Applications* (pp. 610-623). Springer, Cham.

Zarei, K., Farahbakhsh, R., Crespi, N., & Tyson, G. (2020). A first Instagram dataset on COVID-19. *arXiv preprint arXiv:2004.12226*.

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676.

Zarrella, D. Science of Retweets, http://danzarrella.com/the-science-ofretweets-report.html, 2009.

Zhou, J., Zogan, H., Yang, S., Jameel, S., Xu, G., & Chen, F. (2021). Detecting community depression dynamics due to covid-19 pandemic in australia. *IEEE Transactions on Computational Social Systems*.