The Burden of Being a Bridge: Understanding the Role of Multilingual Users during the COVID-19 Pandemic

Ninghan Chen[1], **Zhiqiang Zhong**[1], **Xihui Chen**[2], **Jun Pang**[1,2]
[1] Faculty of Sciences, Technology and Medicine,
University of Luxembourg, L-4364 Esch-sur-Alzette, Luxembourg
[2] Interdisciplinary Centre for Security, Reliability and Trust,
University of Luxembourg, L-4364 Esch-sur-Alzette, Luxembourg

Abstract

The outbreak of the COVID-19 pandemic triggers infodemic over online social networks. It is thus important for governments to ensure their official messages outpace misinformation and efficiently reach the public. Some countries and regions that are currently worst affected by the virus including Europe, South America and India, encounter an additional difficulty: multilingualism. Understanding the specific role of multilingual users in the process of information diffusion is critical to adjust their publishing strategies for the governments of such countries and regions. In this paper, we investigate the role of multilingual users in diffusing information during the COVID-19 pandemic on popular social networks. We collect a large-scale dataset of Twitter from a populated multilingual region from the beginning of the pandemic. With this dataset, we successfully show that multilingual users act as bridges in diffusing COVID-19 related information. We further study the mental health of multilingual users and show that being the bridges, multilingual users tend to be more negative. This is confirmed by a recent psychological study stating that excessive exposure to social media may result in a negative mood.

1 Introduction

The current coronavirus COVID-19 pandemic is a global health crisis of our time. The outbreak of the COVID-19 pandemic leads to an outbreak of information on major online social networks (OSNs), including Twitter, Facebook, Instagram, and YouTube (Cinelli et al. 2020). In this massive COVID-19 outbreak and constantly changing situation, OSNs, thanks to their globally available services, have become essential for people all over the world to seek up-to-the-minute and local information. Moreover, due to the curfew measures and social distancing, people have been spending much more time on OSNs. As a result, social networks have joined conventional media, e.g., TV and radio, to become an indispensable channel for governments and news agents to publish COVID-19 related information. A recent study demonstrated the existence of infodemic on social media during the COVID-19 pandemic and its negative impact on the control of the virus (Cinelli et al. 2020). The term *infodemic* outlines the danger of misinformation during epidemic disease outbreak. As a result, it is very important for governments (and new agents) to design effective strategies to ensure that their official message outpace misinformation and efficiently reach the public.

Some regions that are worst affected by the virus, including European Unions, America and India, have one additional difficulty, that is, their nature of multilingualism. Cross-language channels should be created for people to receive trustworthy information no matter which language they speak. As studied in the literature, multilingual users on social media can work as bridges between language communities (Eleta and Golbeck 2012; Hale 2014). Understanding of the role of multilingual users in information diffusion is thus crucial in practice for governmental departments to design their information release strategy. As far as we know, in spite of the extensive studies explaining and predicting information diffusion (Wang et al. 2017; Chen et al. 2019), it is still unclear whether multilingual users also play a bridging role in information diffusion on social media as they do in social network connections, i.e., whether they act as bridges connecting message originators and message receivers.

The information on OSNs can reflect reality, but the infodemic it triggers can also affect people. Frequent exposure to social media is likely associated with an increase in mental health problems (Gao et al. 2020), including vicarious traumatisation (Liu and Liu 2020), depression (Zhong, Huang, and Liu 2020) and anxiety (Zhong, Huang, and Liu 2020; Amsalem, Dixon, and Neria 2020). Based on these psychological findings, if multilingual users indeed play a critical role in information diffusion, then the resulted excessive exposure to COVID-19 related information may make multilingual users become more stressed or anxious, compared to monolingual users. If this is the case, additional attention should be devoted to multilingual users' mental health.

To sum up, in this paper, we aim to answer the following two research questions:

RQ1: *What role do multilingual users play in the diffusion of COVID-19 related information?*

RQ2: *Are multilingual users more negative than monolingual users during the COVID-19 pandemic? If yes, is the negativity related to their role in information diffusion?*

The information outbreak on social media incurred by the COVID-19 pandemic provides us with valuable data to answer these two questions. We crawled data from Twitter for almost 7 months since the beginning of the COVID-19 pandemic which come from a multilingual region severely hit by the virus: *the Greater Region of Luxembourg* (GR). GR is a cross-border region centred with Luxembourg and composed of its adjacent regions of Belgium, Germany and France. Luxembourg is famous for its multilingualism, as most Luxembourgish people can speak four languages and about 50% of its population are foreigners[1] Luxembourg is severely affected by the virus and ranked in the top three countries in Europe in terms of the number of new infections in every 100,000 inhabitants.

Based on the dataset we collected from Twitter, we examine the bridging role of users in information diffusion by quantifying their influences within information cascades (Chen et al. 2019; Wang et al. 2017). More specifically, we propose two new cascade bridging measures, which allow us to comprehensively understand and quantify how much an OSN user plays the role of as a bridge in COVID-19 related information diffusion. Then we continue to analyse users' mental health status, i.e., being negative or positive, based on the sentiment of their posted tweets during the collection period. In the end, we investigate the correlation between users' role in information diffusion and their mental health.

Our contributions of this paper can be listed as follows:

- We empirically and quantitatively demonstrate the bridging role of multilingual users in the diffusion of COVID-19 related information.
- We analyse users' mental status from the sentiment of their posted messages during the pandemic and discover that multilingual users are more likely to be negative.
- We validate the result of a recent psychological study on social media, revealing the positive correlation of a user's exposure to social media with her/his anxiety. Indeed, we find that users' mental health is strongly correlated with their bridging performance in information diffusion, but has weak or no correlation with their topological properties in the social network.

2 The GR-ego Twitter Dataset

In this section, we describe how we build our GR-ego Twitter dataset, referred to as *GR-ego* dataset in the rest of the paper. In addition to its popularity and large number of active users, we have three other reasons to select Twitter. First, the language of tweets is provided and we can use it to identify multilingual users. If a user only retweets tweets in one single language, we consider her/him as *monolingual*. Otherwise, this user is *multilingual*. Second, the user-input locations of the posters can be used to find users located in GR. Last, it allows us to track the diffusion process of a Twitter message. Specifically, the ID number of the original message is attached if a message is retweeted. We can use it to find other users retweeting the same message and with the retweeting time stamps, we can approximately simulate the diffusion cascade of the message. We give in Table 1 a crawled Twitter message as an example.

Table 1: A GR Twitter message.

Attribute	**Value**
Time	2020-04-28 17:00:09+00:00
Status (Original/Retweeted)	Retweeted
Tweet_id	12319668395******
Full_text	RT @******:The Diamond princess is a UK ship managed by the US. #coronavirus
User_id	u9181074902*****
User_geo_orginal	Moselle
User_geo	Moselle, Lorraine, France
Original_tweet_id	134515******
Original_user_id	103*****
Language	en

Our GR-ego dataset consists of two components: (i) the social network of GR users which records GR users and the following relations between them; (ii) the tweets posted or re-tweeted by GR users during the pandemic. As we will shown in the following sections, these two types of information allow us to capture the diffusion process of COVID-19 related tweets as well as to analyse users' mental sentiment. We perform three steps to collect our GR-ego dataset.

Step 1. Meta data collection. At this step, we aim to collect a set of seed users in GR who are actively involved in COVID-19 discussions.

Due to the poor efficiency to crawl tweets according to keywords, we make use of a public dataset of COVID-19 related tweets (Chen, Lerman, and Ferrara 2020). Restricted by privacy policies, this dataset only consists of tweet IDs. With these IDs, we subsequently crawl their content through the Twitter API. We extract the messages posted in the period between the early stage of the pandemic (January 22nd, 2020) and the end of the first wave of the pandemic (mid July, 2020), for about 7 months. In Twitter, geography information, e.g., the locations of the posters and the original users if messages are re-tweeted, is provided by Twitter users themselves. As a result, they are usually ambiguous and do not have a unified format. We leverage the geocoding APIs, Geopy[2] and ArcGis Geocoding[3] to regularise the location format and remove the ambiguity. For instance, the location of the poster of the example message in Table 1, i.e., *Moselle*, is transformed to a preciser and machine-parsable address: *Mosselle, Lorraine, France*. With the transformed geo-locations, we collect the Twitter messages from the greater region. In the end, we obtain 128,310 tweets in total from 8,872 GR users.

Step 2. Social network construction. At this step, our purpose is to search more GR users from the seed users and construct the GR-ego social network.

We use an iterative approach to gradually enrich the social graph until it stops growing. We start with the seed users. For each user, we obtain their followers and only keep the ones that have a mutual following relation with the seed user. The

[1] https://bit.ly/3fdhgwj

[2] https://bit.ly/3gfW2PP

[3] https://bit.ly/3f9OUDa

reason is that such users have much larger probabilities to reside in GR, which ensures a good efficiency of construction. The locations of the new users are obtained through their posts and processed in the same way as in the previous step. Only users from GR are added to the social network as new nodes together with their following relations with other existing users. Specifically, if user u follows u', a directed edge from node u' to u will be added.

After the first round, we continue going through the newly added users one after another by adding their mutually followed friends that do not exist in the current social network. This process will continue until no new users are added into the network. In our case, it takes 5 iterations before termination. We take the largest weakly connected component of the constructed social network as the *GR-ego social network*. Table 2 summarises its main statistics.

#node	5,808,938
#edge	12,511,698
Average degree	2.15
Average local clustering coefficient	0.04

Table 2: Statistics of the GR-ego social network.

Step 3. Timeline tweets crawling. At this step, we collect tweets posted or re-tweeted *during the pandemic* by GR users to construct the dataset of Twitter messages. Recall that these past tweets will be used to analyse the mental status of the posters during the pandemic.

According to the Twitter policy, we can only download the last 3,200 tweets in a user's timeline. Even with this limit, it is still rather time-consuming to download the tweets of all the users in the GR-ego social network constructed in the previous step. Therefore, we select a subset of representative users with a sufficiently large size and crawl their tweets. Specifically, we choose the 8,872 users in the meta dataset together with their followers and followees, which add up to 15,255 users. From these users, we collect 10,994,189 Twitter messages in total. Figure 1 presents the numbers of the tweets in different languages. We can see that the collected tweets are posted in a very diverse set of languages. English, as a universal language, is still the dominant language and the distribution of other languages are consistent with that of the nationalities of the GR inhabitants.

3 Data Processing

We conduct two processing steps to obtain information required for our analysis.

3.1 Cascade computation

A cascade records the process of the diffusion of a message. It stores all activated users (in our dataset, users retweeting the message) and when they are activated, e.g., the order of activation. In this paper, we adopt a widely accepted model to represent the cascade of a message: *cascade tree* (Wang et al. 2017).

The user who first posts the message is the root of a cascade tree. The users who retweet the message but have no followers retweeting the message comprise the leaf nodes. An edge from u to u' is added to the cascade if u' follows u and u' re-tweeted the message after u. If many users who u' follows retweeted the message, then we select the last one retweeting the message as the parent node of u'. Figure 2(b) shows a cascade example for the social network as depicted in Figure 2(a). In this example, user u_4 can be activated by the messages retweeted by either u_1 or u_3. Since u_3 retweeted after u_1, we add the edge from u_3 to u_4 to indicate that the retweeting of u_3 has activated u_4.

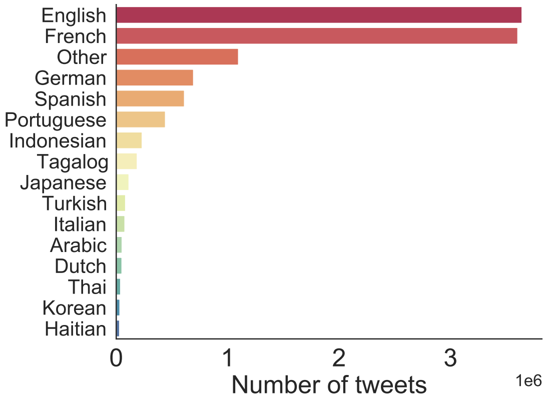
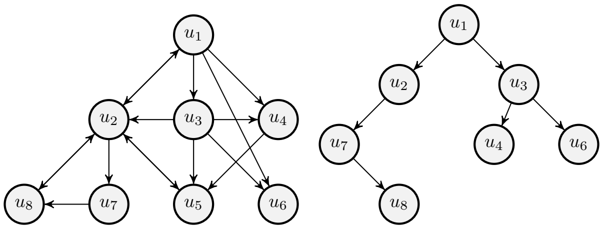


Figure 1: Distribution of different languages of GR Twitter messages.



(a) A social graph (b) A cascade

Figure 2: Example of a cascade.

We denote the root node of a cascade C by $r(C)$. We call a path that connects the root and a leaf node a *cascade path*, which is actually a sequence of nodes ordered by their activation time. For instance, (u_1, u_3, u_4) is a cascade path in our example indicating that the diffusion of a message started from u_1 and reached u_4 in the end through u_3. In the rest of this paper, we represent a cascade tree as a set of cascade paths for the purpose of simplicity. For instance, the cascade in Figure 2(b) is denoted by the following set $\{(u_1, u_2, u_7, u_8), (u_1, u_3, u_4), (u_1, u_3, u_6)\}$.

We follow the method in (Kupavskii et al. 2012; Tsur and Rappoport 2012) to construct cascades. Recall that when a tweet's status is 'Retweeted', the ID number of the original tweet is attached to the retweeted message (see in Table 1). We first create a set of original tweets consisting of the tweets in our meta data with the status as 'Original'. Second, for each original tweet, we collect the user IDs that have retweeted the message. At last, we generate the cascade based on the following relationship in our GR-ego social network and their retweeting time stamps. We eliminate cascades with only two users where a message is just retweeted once. Table 3 summarises the statistics of the collected cascades. In total, 29,710 cascades are built with $83,904$ users involved. An interesting observation is that multilingual users are very active in diffusing COVID-19 related information. Among all the participants, only 5.04% are multilingual users but they participate in 68.5% of the cascades. On average, each multilingual user diffuses 13.11 messages, which is 2 times more than that of monolingual users. Another observation is that the cascades with multilingual users have 2 more users on average.

Table 3: Cascade statistics.

#cascade	29,710
%cascade with multilingual users	68.5%
#cascade per multilingual user	13.11
#cascade per monolingual user	5.69
#participant	83,900
%multilingual participants	5.04%
#user per cascade (with multilingual users)	7.2
#user per cascade (only monolingual users)	5.2

3.2 Sentiment analysis

Previous works (Zhou, Jin, and Zafarani 2020) leverage user-provided mood (e.g., angry, excited) or status to analyse users' mental status. However, such information is not available in most popular social networks (e.g., Twitter). Fortunately, with the recent fast advance of sentiment analysis methods (Balahur and Turchi 2012; Devlin et al. 2018), we can still obtain users' sentiment by analysing their posted Twitter messages. Sentiment analysis for multilingual text faces a serious lack of resources (Balahur and Turchi 2012). Most works choose to translate the original language into a resource-rich language, such as English, and then analyse with existing lexicons or annotated dataset (Denecke 2008). Transformer-based representation learning methods (Devlin et al. 2018) allow us to obtain users' sentiment by analysing their posted tweet without translation. In this paper, we make use of the methods for polarity sentiment analysis. In other words, we only classify the sentiment of a message as negative or positive.

We build an end-to-end deep learning model to conduct the classification which is composed of three components. As the first component, we use XLM-RoBERTa (Ou and Li 2020), a pre-trained multilingual language model, to obtain the embedding of tweet content text. The embedding results are feed into our second component, a fully-connected ReLU layer with dropout. At last, we add a linear layer on top of the pre-trained model's outputs for regression with an activation function, i.e., sigmoid. We use cross-entropy as our loss function and optimise it with the Adam optimizer in our representation learning process.

We train our model on the Sentiment140 dataset (Go, Bhayani, and Huang 2009). The dataset contains 1.6 million tweets collected with keywords. All tweets are annotated with a binary sentiment value indicating positive or negative. We split the dataset and take the first 80% of the tweets as the training set and the rest 20% for testing. We assign other training parameters following the common principle used in existing works. Specifically, we run 10 epochs with the maximum string length set to 128 and dropout ratio to 0.5. When tested with Micro-F1 score and accuracy metrics, we get an accuracy of 85.48% and F1 Score 84.11%.

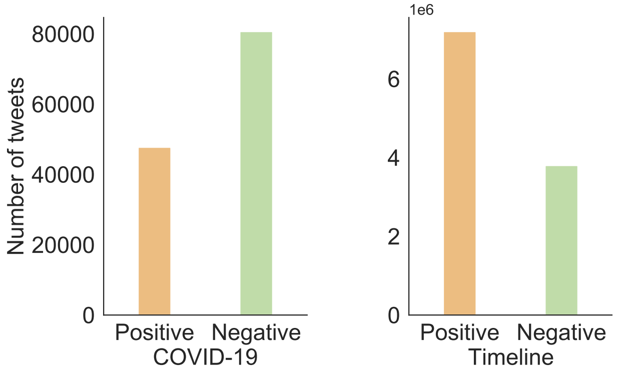


Figure 3: Sentiment distribution of diffused COVID-19 related information (left) and users' timeline tweets (right): Negative COVID-19 information is diffused while users tend to be positive.

Before applying our sentiment classification model, we clean the tweet contents by removing all URLs, mentioned usernames and the word 'RT'. Figure 3 summarises the results for the COVID-19 related tweets and the user timeline tweets. The results from users' timeline tweets are consistent with the existing work (Pak and Paroubek 2010), i.e., users tend to post positive messages in online social media. It is also reasonable that 62.8% COVID-19 related tweets are negative considering the continuous large-scale infections.

4 Bridging Role of Multilingual Users in Information Diffusion

In this section, we will answer the first research question (RQ1). Specifically, we will quantitatively analyse the role of multilingual users in the diffusion of COVID-19 related information. We will check whether multilingual users act as bridges between the originator and activated users. We propose two new measures to quantify multilingual users' importance in information diffusion from two different levels. The first one is defined from the level of users and evaluates an individual user's overall performance in COVID-19 information diffusion. It allows us to compare users in terms of their overall bridging importance. The second one focuses on the level of cascades. It compares the bridging performance of multilingual users and monolingual users, as two separate groups, for each cascade. It can be used verify whether the multilingual group outperforms the other group.

4.1 User information diffusion importance

We evaluate each user's overall performance in the diffusion of all observed Twitter messages. As a user can participate in the diffusion of many messages, before calculating the user's overall performance, we start with her/his importance on the diffusion of one single message and then combine her/his importance of all messages into one single measurement.

Intuitively, we consider a user as more important in the diffusion of a message according to three criteria:

1. the user participates a larger number of cascade paths;
2. the cascade paths with the users have larger lengths;
3. more users get activated directly from him.

The first two criteria assign more importance to users who can activate more users with her/his information sharing behaviour. It has been pointed out in the literature (Zhou et al. 2010) that cascades tend to be wide but not deep and wider cascades can facilitate the receipt of the message by more diverse people. In the scenario of multilingualism, more users activated directly by a multilingual user will imply a larger chance of diffusion across different language communities.

Given a cascade path $S = (u_1, u_2, \dots, u_n)$, we use $S^*(u_i)$ $(1 \leq i < n)$ to denote the subsequence composed of the nodes after u_i (including u_i), i.e., $(u_i, u_{i+1}, \dots, u_n)$. For any u that does not exist in S, we have $S^*(u) = \varepsilon$ where ε represents an empty sequence and its length$|\varepsilon| = 0$.

Definition 1 (Cascade bridging value) *Given a cascade tree C and a user u ($u \neq r(C)$), the cascade bridging value of u in C is calculated as:*

$$\alpha_C(u) = \frac{\sum_{S \in C} \frac{|S^*(u)|}{|S|}}{|C|}$$

Note that our purpose is to evaluate the importance of users being a transmitter or sharer of a message. Therefore, the concept of cascade bridging value is not applicable to the root user, i.e., the message originator.

Example 1 *We take user u_3 in Figure 2(b) as an example. He participates two out of the three cascade paths, i.e., $S_1 = (u_1, u_3, u_4)$ and $S_2 = (u_1, u_3, u_6)$. Thus $S_1^* = (u_3, u_4)$ and $S_2^* = (u_3, u_6)$. We then have $\alpha_C(u_3) = \frac{2/3+2/3}{3} \approx 0.44$.*

Note that we do not simply use the proportion of users that get activated from a user in a cascade to evaluate her/his bridging role. This is because it cannot distinguish users with directly activated followers. Taking u_2 in Figure 2(b) as an example, according to our definition, $\alpha_C(u_2) = 0.25$ is smaller than $\alpha_C(u_3)$ because u_3 directly activated two users. However, if we take the proportion of activated users, the values of these two users will be the same.

With a user's bridging value calculated in each cascade, we define *user bridging magnitude* to evaluate her/his overall importance in the diffusion of a given set of observed messages. Intuitively, we first add up the bridging values of a user in all his/her participated cascades and then normalise the sum by the maximum number of cascades participated by a user. It captures not only the bridging value of a user in each participated cascade, but also the number of cascades he participated in. This means, a user who is more active in sharing COVID-19 related information is considered more important in information diffusion.

Definition 2 (User bridging magnitude (UBM)) *Let $\mathcal{C}$ be a set of cascades on a social network and $\mathcal{U}$ be the set of users that participate at least one cascade in $\mathcal{C}$. A user u's information diffusion importance is calculated as the average cascade bridging value over the cascades in $\mathcal{C}$, i.e.,*

$$\omega_{\mathcal{C}}(u) = \frac{\sum_{C \in \mathcal{C}} \alpha_C(u)}{\max_{u' \in \mathcal{U}} |\{C \in \mathcal{C} | \alpha_C(u') > 0\}|}.$$

With this measure, for any two users, we can compare their UBM values and learn which one plays a more important role in information diffusion.

Empirical verification. We analyse and compare the overall importance of multilingual and monolingual users in information diffusion during the pandemic based on their UBM values. If multilingual users play a more important bridging role, we should have i) multilingual users normally have larger UBM values; and ii) a larger proportion of multilingual users should have good UBM values than monolingual users. We will verify whether these two expectations can be observed in our GR-ego dataset. Figure 4 shows the UBM distribution of the two groups of users from three perspectives. The first general observation is that over 97% of users' UBM values lie in the range between 0.0 and 0.4. We consider UBM that is smaller than 0.1 as *weak* and that larger than 0.2 as *strong*.

From the box plots on the left, we can see that the median and mean (labelled by green triangles in the boxes) of multilingual users' UBM values are over 2 times as big as those of monolingual users. This indicates that in general multilingual users have larger UBM values.

We show the complementary cumulative density function (CDDF) in the middle and the probability density function (PDF) of UBM on the right. We have two observations. First, for a significantly large proportion of monolingual users (approximately 78%), their UBM values are weak, i.e., smaller than 0.1. By contrast, only 46% multilingual users have weak UBM values. This indicates that a large number of monolingual users cannot effectively activate other users to diffuse messages. Second, it is clear that compared to monolingual users, more multilingual users have large UBM values. About 23% of multilingual users hold strong UBMs which is over 3 times more than monolingual users.

From the above analysis, we can see that our two expectations are observed in our dataset. Therefore, we conclude that multilingual users can activate more users in diffusing COVID-19 related messages.

4.2 Cascade bridging magnitude

Our first measure focuses on the level of users and evaluates their overall performance across all observed information cascades. It does not consider the relative bridging performance of multilingual users and monolingual users within the cascades. We take an example to explain this.

Example 2 *We still take the cascade in Figure 2(b) as an example. Suppose u_2 is the only multilingual user. After calculating the cascade bridging values, we can learn that $\alpha_C(u_2) = 0.25$, and as a multilingual user, u_2 plays a more important role in diffusing the message than all monolingual*

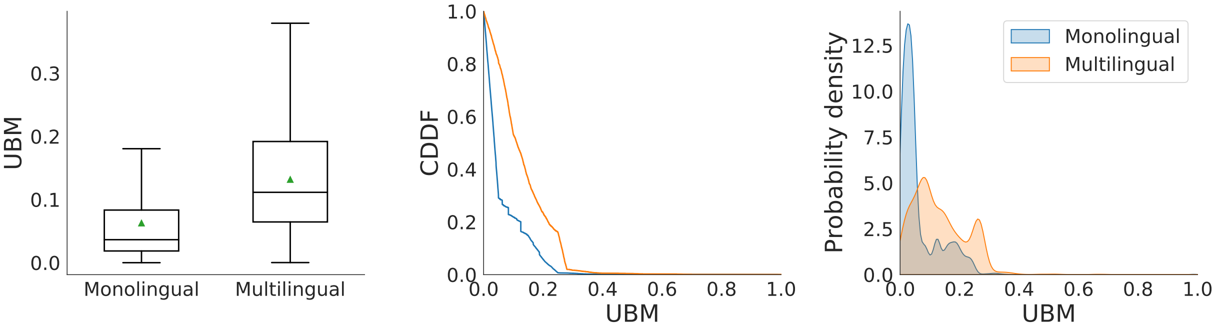


Figure 4: UBM distributions. The box plot (left) shows that multilingual users in general have larger UBM values. The CDDF (middle) and PDF (right) demonstrate together that a larger proportion of multilingual users have good UBM values.

users except for u_3 with $\alpha_C(u_3) = 0.44$. In this example, we still cannot say multilingual users as a whole play a more important bridging role in this cascade.

Therefore, we propose a second measure to compare the performance of multilingual users to monolingual users.

Given a cascade C with multilingual users, we use $\mathcal{U}_C^{mul}$ to denote the set of multilingual users and $\mathcal{U}_C^{mon}$ the set of monolingual users in the cascade. Then we calculate an integrated value through a function γ from the bridging values of multilingual users and that of the monolingual users, represented by α_C^{mul} and α_C^{mon}, respectively. The integrated value can be the *mean*, *median* or *maximum*. Formally,

$$\begin{aligned} \alpha_C^{mul} &= \gamma(\{\alpha_C(u) \mid u \in \mathcal{U}_C^{mul}\}), \\ \alpha_C^{mon} &= \gamma(\{\alpha_C(u) \mid u \in \mathcal{U}_C^{mon}\}) \end{aligned}$$

Note that the integration function γ should be instantiated according to practical requirements. We say multilingual users play a bridging role in a cascade C when $\alpha_C^{mul} > \alpha_C^{mon}$. In the end, we use the notion *cascade bridging magnitude* (CBM) to quantify the importance of multilingual users as a whole in information diffusion.

Definition 3 (Cascade bridging magnitude (CBM)) *Let $\mathcal{C}_{mul} \subseteq \mathcal{C}$ be the set of cascades involving at least one multilingual user. Then the bridging magnitude is calculated as the following:*

$$M = \frac{\sum_{C \in \mathcal{C}_{mul}} \mathbb{1}(\alpha_C^{mul} > \alpha_C^{mon})}{|\mathcal{C}_{mul}|}$$

where $\mathbb{1}(x < y)$ is an indicator function which returns 1 *when $x < y$ and 0, otherwise.*

Experimental evaluation. In Table 4, we list the results about the CBM values calculated with our GR-ego dataset. In our analysis, we instantiate the integration function γ with *maximum*, *media* and *mean*.

Instead of just showing the cascade bridging magnitude of multilingual users, we also present the statistics in another two cases. Specifically, we use the term 'hold' to indicate the case when multilingual users play a more important bridging role in the cascade (i.e., $\alpha_C^{mul} > \alpha_C^{mon}$), while 'not hold' indicates the opposite case. When multilingual users and monolingual users have the same integrated bridging value, it is 'uncertain' who are more important in the cascade. Before analysing the statistics, we will set a criterion that if multilingual users performs better in more than 50% cascades with multilingual participants, we will say the bridging role of multilingual users in the level of cascades strongly holds. If it is between 30% and 50%, considering the comparably small number, we say the bridging role weakly hold. Otherwise, the bridging role does not hold.

Table 4: The cascade bridging magnitude of multilingual users. Multilingual users have performed dominantly better in all the three integration functions.

Operation		#cascade	%cascade
Maximum	**hold**	18,908	63.64%
	not hold	9,545	32.13%
	uncertain	1,257	4.23%
Median	**hold**	18,903	63.63%
	not hold	9,540	32.11%
	uncertain	1,267	4.26%
Mean	**hold**	18,914	63.66%
	not hold	9,557	32.17%
	uncertain	1,239	4.17%

The obvious observation from Table 4 is that multilingual users has a cascade bridging magnitude of 63.64% on average under all the three integration functions (i.e., maximum, mean and median). This means that multilingual users play a more important bridging role in about 65% cascades, which almost doubles that of monolingual users. Considering the small percentage of multilingual users in all the participants (i.e., 5.04%), we can conclude that multilingual users' bridging role in COVID-19 related information diffusion in the level of individual cascades strongly holds.

Summary of the section. From the above discussion, we can see that multilingual users perform dominantly better at both the user level and the cascade level. Therefore, we conclude that multilingual users play an important bridging role in COVID-19 related information diffusion.

5 The Burden of Being a Bridge

In the previous section, we have succeeded in demonstrating that multilingual users play a bridging role in diffusing

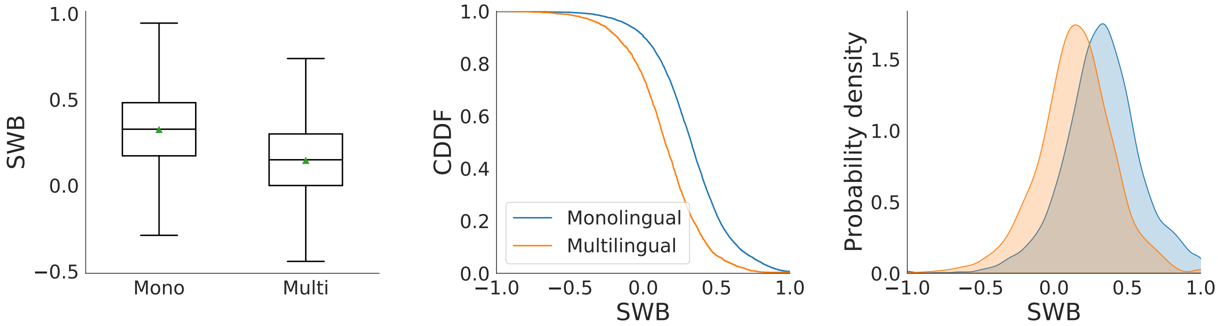


Figure 5: Distribution of SWB values. The box plot (left) shows that multilingual users are about 2 times more negative. The CDDF (middle) and PDF (right) distributions demonstrate that a much smaller proportion of multilingual users have a relatively good SWB value.

COVID-19 related information. In this section, we proceed with answering our second research question (RQ2), i.e., to check whether multilingual users are more negative and the correlation with their bridging role.

5.1 Multilingualism results in negativity

As we mentioned previously, we analyse the mental well-beings of users based on the sentiments of their posted messages and verify whether multilingual users are more negative. In Section 3, we have obtained the sentiments of users' timeline messages. We make use of the notion of *subjective well-being score* (SWB) proposed in (Zhou, Jin, and Zafarani 2020) to quantify the extent of mental positivity of a user based on their past posts.

Definition 4 (Subjective well-being score (SWB)) *We use $N_p(u)$ and $N_n(u)$ to denote the number of positive posts and the number of negative posts of a user u, respectively. The subjective well-being score of u, represented by $swb(u)$, is calculated as:*

$$swb(u) = \frac{N_p(u) - N_n(u)}{N_p(u) + N_n(u)}$$

A larger SWB value indicates the corresponding user is more positive.

Empirical verification. We proceed to compare the mental well-beings of the two groups of users. We say a user is *consistently positive* if her/his SWB value is larger than 0.5 and a user is *consistently negative* if her/his SWB value is smaller than 0.0. We select the value of 0.5 as only 20% users in our dataset have larger SWB values.

Similar to the previously analysis, we first give our criteria that will be used to determine whether multilingual users tend to be more negative. Three conditions should be simultaneously satisfied. First, in general, multilingual users are more likely to be negative. Second, a larger proportion of multilingual users have consistent negative mental status. Third, a smaller proportion of multilingual users are consistently positive.

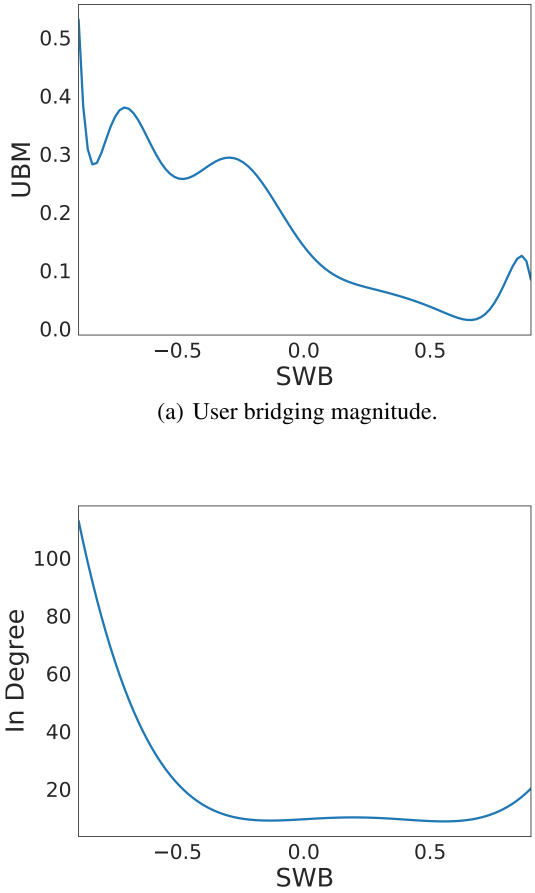
Figure 5 presents the distributions of the SWB values of multilingual and monolingual users. We now go through the three conditions one after another. From the plot box on the left, we can clearly see that the mean and median SWB value of multilingual users are over two times smaller. This implies that multilingual users are generally more negative than monolingual users. From the rest two distributions, we can get about 20% multilingual users are consistently negative while only 8% monolingual users are consistently negative. Furthermore, about 5% multilingual users are consistently positive which is only one fourth of that of monolingual users. Therefore, the last two conditions are satisfied.

From the above analysis, we conclude that multilingual users are indeed more negative than monolingual users.

5.2 Correlation to users' bridging role

We continue to understand whether the comparatively larger negativity of multilingual users actually results from their dual bridging role in *social network connections* and *information diffusion*. To achieve this goal, we verify whether a user's sentiment is correlated to their bridging role. In terms of global network connections, we take the in-degree and out-degree of nodes, i.e. the number of followees and followers, to represent users' topology features, as commonly used in the literature (Zhou et al. 2010; Agarwal et al. 2020). With respect to the bridging role in information diffusion, we use the measure *user bridging magnitude* as we focus on the correlation of a user's bridging performance with her/his own mental sentiment.

In Table 5, we show the correlation coefficients calculated for the three selected features. It is clear that the out-degrees have no relation with users' SWB values due to the small correlation coefficient 0.008. With the correlation coefficient of -0.69, we can interpret that users' information diffusion bridging performance quantified by our proposed measure is strongly correlated to their mental well-being. According to the correlation coefficient of 0.08, the in-degrees have a weak correlation with SWBs. We visualise the correlation of UBM values and in-degrees with SWB values in Figure 6 for the purpose of cross-checking. As we discussed previously, 97% of users' SWB values mainly span between 0 and 0.4. Thus, we can only concentrate the part lying in this range. We can see that UBM decreases almost linearly when SWB increases while the in-degree remains unchanged. This means in-degrees actually do not correlate to SWB values.



(a) User bridging magnitude.

(b) In-degree.

Figure 6: Correlation between users' sentiment and their UBM/in-dgree.

In summary, we conclude that a user's mental sentiment during the COVID-19 pandemic is strongly correlated to her/his user bridging magnitude but is not correlated with their topology properties in the social network, e.g., in-degree and out-degrees.

Table 5: Correlations between users' SWB and their UBM and structural properties (in-degree and out-degree).

	Correlation Coefficient
[SWB, user bridging magnitude]	-0.69 ($p < 0.005$)
[SWB, in-degree (#followees)]	0.49 ($p < 0.005$)
[SWB, out-degree (#followers)]	0.08 ($p < 0.005$)

6 Related Work

The role of being undisputed universal lingua franca of information diffusion on a global level for English has been questioned in the COVID-19 outbreak (Piller, Zhang, and Li 2020). It has been found that on Twitter nearly 49% of posts are written in a language different from English (Hong, Convertino, and Chi 2011). With their ability to cross the language barrier on social networks, bilingual and multilingual users have attracted special attention from researchers (Eleta and Golbeck 2014, 2012; Kim et al. 2014) as well.

Hale (Hale 2014) studied multilingual users from the view of the topological structure of social networks and found that without multilingual users, social networks will be disconnected. This implies multilingual users play a bridging role in the connectivity of the network. Eleta et al. (Eleta and Golbeck 2012) discovered that multilingual users also act as bridges between communities speaking different languages. With regard to information diffusion, it is studied that non-native English speakers have higher influence than native English users (Kim et al. 2014). Agarwal et al.(Agarwal et al. 2020) showed that multilingual users play a special role in cross-lingual diffusion and multilingual users select the language for their tweet according to audiences (Johnson 2013; Murthy et al. 2015; Nguyen, Trieschnigg, and Cornips 2015). Different from the existing works in the literature, in this paper, we investigate multilingual users' importance in diffusing messages, namely whether messages can reach more user due to the participation of multilingual users.

7 Conclusion

In this paper, we have successfully answered two research questions. The first question is what role multilingual users play in COVID-19 information diffusion. The second is whether multilingual users' mental health will be negatively affected by their role in information diffusion during the pandemic. In the following, we briefly describe our findings and discuss their potential value in practice.

For the first question, through the Twitter dataset we collected during the COVID-19 pandemic, we have empirically shown multilingual users have been playing an important bridging role in diffusing COVID-19 related information. Thanks to their active participation and influence, COVID-19 information can be spread to more users. Our finding is of great value in practice for the multilingual countries and regions to fight against infodemic and mitigate its damage. Official messages should be released in a carefully designed manner so that a large number of multilingual users could be activated in the early stage of the diffusion process. An interesting piece of future work along this direction is thus to study optimal approaches to activate multilingual users as many as possible within a given time window.

For the second question, we discovered multilingual users are at higher risk of being affected by infodemic compared to monolingual users due to their repeated exposure to COVID-19-related information (Holmes et al. 2020). At the time of writing, the pandemic is still evolving. We believe that this could draw special public attention on the mental health on multilingual users and may incur studies in the multilingual countries and regions that are badly hit by the virus on new approaches to mitigate this potential mental health crisis.

In this paper, our study is based on the data during the COVID-19 pandemic. One of our future works is thus to check whether our findings in this paper also hold in the general information diffusion process.

References

Agarwal, P.; Garimella, K.; Joglekar, S.; Sastry, N.; and Tyson, G. 2020. Characterising user content on a multilingual social network. In *Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM)*, volume 14, 2–11.

Amsalem, D.; Dixon, L. B.; and Neria, Y. 2020. The coronavirus disease 2019 (COVID-19) outbreak and mental health: current risks and recommended actions. *JAMA Psychiatry* .

Balahur, A.; and Turchi, M. 2012. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, 52–60.

Chen, E.; Lerman, K.; and Ferrara, E. 2020. COVID-19: The first public coronavirus Twitter dataset. *arXiv preprint arXiv:2003.07372* .

Chen, X.; Zhang, K.; Zhou, F.; Trajcevski, G.; Zhong, T.; and Zhang, F. 2019. Information Cascades Modeling via Deep Multi-Task Learning. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 885–888. ACM.

Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C. M.; Brugnoli, E.; Schmidt, A. L.; Zola, P.; Zollo, F.; and Scala, A. 2020. The COVID-19 social media infodemic. *Scientific Reports* 10(1): 16598.

Denecke, K. 2008. Using sentiwordnet for multilingual sentiment analysis. In *Proceedings of the Workshops at the 24th International Conference on Data Engineering (ICDE)*, 507–512.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Eleta, I.; and Golbeck, J. 2012. Bridging languages in social networks: How multilingual users of Twitter connect language communities? *Proceedings of the American Society for Information Science and Technology* 49(1): 1–4.

Eleta, I.; and Golbeck, J. 2014. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior* 41: 424–432.

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

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* 1(12): 2009.

Hale, S. A. 2014. Global connectivity and multilinguals in the Twitter network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 833–842.

Holmes, E. A.; O'Connor, R. C.; Perry, V. H.; Tracey, I.; Wessely, S.; Arseneault, L.; Ballard, C.; Christensen, H.; Silver, R. C.; Everall, I.; et al. 2020. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. *The Lancet Psychiatry* .

Hong, L.; Convertino, G.; and Chi, E. 2011. Language matters in twitter: A large scale study. In *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM)*, volume 5.

Johnson, I. 2013. Audience design and communication accommodation theory: Use of Twitter by Welsh-English biliterates. *Social Media and Minority Languages: Convergence and the Creative Industries* 99–118.

Kim, S.; Weber, I.; Wei, L.; and Oh, A. 2014. Sociolinguistic analysis of Twitter in multilingual societies. In *Proceedings of the 25th ACM conference on Hypertext and Social Media*, 243–248.

Kupavskii, A.; Ostroumova, L.; Umnov, A.; Usachev, S.; Serdyukov, P.; Gusev, G.; and Kustarev, A. 2012. Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM)*, 2335–2338.

Liu, C.; and Liu, Y. 2020. Media exposure and anxiety during COVID-19: The mediation effect of media vicarious traumatization. *International Journal of Environmental Research and Public Health* 17(13): 4720.

Murthy, D.; Bowman, S.; Gross, A. J.; and McGarry, M. 2015. Do we tweet differently from our mobile devices? a study of language differences on mobile and web-based twitter platforms. *Journal of Communication* 65(5): 816–837.

Nguyen, D.; Trieschnigg, D.; and Cornips, L. 2015. Audience and the use of minority languages on Twitter. In *Proceedings of the 9th nternational AAAI Conference on Web and Social Media (ICWSM)*, volume 9.

Ou, X.; and Li, H. 2020. YNU_OXZ @ HaSpeeDe 2 and AMI : XLM-RoBERTa with Ordered Neurons LSTM for Classification Task at EVALITA 2020. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian.*

Pak, A.; and Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC*. European Language Resources Association.

Piller, I.; Zhang, J.; and Li, J. 2020. Linguistic diversity in a time of crisis: Language challenges of the COVID-19 pandemic. *Multilingua* 39(5): 503–515.

Tsur, O.; and Rappoport, A. 2012. What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM)*, 643–652.

Wang, Y.; Shen, H.; Liu, S.; Gao, J.; and Cheng, X. 2017. Cascade Dynamics Modeling with Attention-based Recurrent Neural Network. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2985–2991. ijcai.org.

Zhong, B.; Huang, Y.; and Liu, Q. 2020. Mental health toll from the coronavirus: Social media usage reveals Wuhan residents' depression and secondary trauma in the COVID-19 outbreak. *Computers in Human Behavior* 114: 106524.

Zhou, X.; Jin, S.; and Zafarani, R. 2020. Sentiment Paradoxes in Social Networks: Why Your Friends Are More Positive Than You? In *Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM)*, 798–807. AAAI Press.

Zhou, Z.; Bandari, R.; Kong, J. S.; Qian, H.; and Roychowdhury, V. P. 2010. Information resonance on Twitter: watching Iran. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis (SNAKDD)*, 123–131. ACM.