Validating part of the social media infodemic listening conceptual framework using structural equation modelling

*Shu-Feng Tsao, Helen Chen, and Zahid A. Butt**

School of Public Health Sciences, Faculty of Health, University of Waterloo, Waterloo, Ontario, Canada

eClinicalMedicine
2024;70: 102544

Published Online xxx
https://doi.org/10.1016/j.eclinm.2024.102544

Summary

Background The literature has identified various factors that promote or hinder people's intentions towards COVID-19 vaccination, and structural equation modelling (SEM) is a common approach to validate these associations. We propose a conceptual framework called social media infodemic listening (SoMeIL) for public health behaviours. Hypothesizing parameters retrieved from social media platforms can be used to infer people's intentions towards vaccination behaviours. This study preliminarily validates several components of the SoMeIL conceptual framework using SEM and Twitter data and examines the feasibility of using Twitter data in SEM research.

Methods A total of 2420 English tweets in Toronto or Ottawa, Ontario, Canada, were collected from March 8 to June 30, 2021. Confirmatory factor analysis and SEM were applied to validate the SoMeIL conceptual framework in this cross-sectional study.

Findings The results showed that sentiment scores, the log-numbers of favourites and retweets of a tweet, and the log-numbers of a user's favourites, followers, and public lists had significant direct associations with COVID-19 vaccination intention. The sentiment score of a tweet had the strongest relationship, whereas a user's number of followers had the weakest relationship with the intention of COVID-19 vaccine uptake.

Interpretation The findings preliminarily validate several components of the SoMeIL conceptual framework by testing associations between self-reported COVID-19 vaccination intention and sentiment scores and the log-numbers of a tweet's favourites and retweets as well as users' favourites, followers, and public lists. This study also demonstrates the feasibility of using Twitter data in SEM research. Importantly, this study preliminarily validates the use of these six components as online reaction behaviours in the SoMeIL framework to infer the self-reported COVID-19 vaccination intentions of Canadian Twitter users in two cities.

Funding This study was supported by the 2023-24 Ontario Graduate Scholarship.

Keywords: Structural equation modelling; Twitter; COVID-19; Vaccine intention

Introduction

Throughout the COVID-19 pandemic, social media have played a substantial role in shaping public perceptions and attitudes towards COVID-19 vaccination.[1,2] As a result of extreme interventions such as lockdowns to contain COVID-19 transmission before vaccines were available, people have increasingly connected and relied on digital channels, such as social media, to receive information related to COVID-19. Although social media platforms can be useful tools for disseminating accurate and helpful information, they also fuel vaccine hesitancy.[1–6] The spread of misinformation about COVID-19 vaccines has been a breeding ground for vaccine hesitancy given conspiracy theories and other misleading information regarding vaccine safety and efficacy, polarization, and emotions, which can easily go viral and create doubts among users.[1–6] In addition to typical online questionnaires or qualitative analysis, researchers have applied machine learning (ML) or artificial intelligence (AI) techniques to investigate and better understand public discourse and sentiments and infer people's COVID-19 vaccine intentions.[7–9] The World Health Organization coined the term "social listening" to describe such activities and deployed its Early AI-Supported Response with Social Listening (EARS) platform during the pandemic.[10]

Social listening studies have adopted existing theories from health behaviours, communication, and

*Corresponding author.
E-mail address: zahid.butt@uwaterloo.ca (Z.A. Butt).

Research in context

Evidence before this study

Typical structural equation modelling (SEM) research has used online surveys to examine people's intentions to accept COVID-19 vaccines during the pandemic. Various theories, such as the health belief model and the theory of planned behaviour, have been commonly adopted in previous research. However, existing theories have limitations given the complex information ecosystems in modern societies, especially social media. We propose social media infodemic listening (SoMeIL) as a conceptual framework for public health behaviours. However, validation of the proposed conceptual framework is needed. Since this framework is developed according to social media, it requires the use of social media data. However, social media data, such as Twitter data, have rarely been used in SEM research, although they have been analysed in other studies that have investigated people's intentions or behaviours in relation to the COVID-19 vaccination.

Added value of this study

The findings of this study indicate significant statistical relationships between COVID-19 vaccination intention and several components derived from Twitter, including a tweet's sentiment score, the numbers of a tweet's favourites and retweets, and the numbers of Twitter users' favourites, followers, and public lists. Therefore, this study provides a preliminary validation of the proposed SoMeIL conceptual framework. This study also demonstrates the feasibility of using Twitter data rather than survey data in SEM research. For public health contexts, some indicators on Twitter, such as the numbers of likes and shares, can be used to infer Canadian Twitter users' vaccination behaviours in real life. Therefore, the findings of this study can be adopted and expanded to forecast vaccination coverage for vaccine-preventable diseases. This approach can also help to tailor communication strategies and address specific issues based on Twitter users' discussions and online behaviours to effectively reach different groups. The SoMeIL conceptual framework can be extended to other areas, such as symptom reports or behavioural patterns, to aid in public health decision-making and resource allocation. By integrating social media platforms such as Twitter into pandemic preparedness, health organizations and government agencies can harness their potential as powerful tools to engage with the public, address health misinformation, and effectively respond to crises, ultimately helping to mitigate the impact of future pandemics. Similar to other pandemic surveillance platforms, the SoMeIL conceptual framework can provide real-time monitoring and surveillance since social media data can complement traditional surveillance methods and help public health authorities respond quickly to potential outbreaks.

Implications of all the available evidence

Our findings preliminarily validate the proposed conceptual framework and show that social media data from Twitter can be used in SEM research. The best model demonstrates that the four variables derived from Twitter can be used as proxies linked to Canadian Twitter users' intentions to receive the COVID-19 vaccine. However, additional studies are needed to further confirm the proposed conceptual framework with different model specifications and social media data.

behavioural sciences.[7–10] However, these models have limitations, and they do not truly reflect current complex information ecosystems. The literature includes various social listening studies that investigate the impacts of exposure to information circulated on social media platforms on individuals' intentions or behaviours in relation to COVID-19 vaccination.[11] This research has generally been conducted using surveys and statistical analyses, such as structural equation modelling (SEM), to identify associations.[12–17] SEM has been widely used to investigate factors that influence people's intentions or attitudes towards the COVID-19 vaccination through the use of different theories as well as latent and multiple dependent variables.[12–17] A distinctive feature of SEM is its incorporation of latent variables, which are theoretical constructs that cannot be directly observed or measured but are inferred from a set of measured indicators, such as intentions or perceptions.[12–17] SEM allows researchers to investigate complex relationships among variables, both observable and latent, and offers a practical and flexible tool for understanding complex structural associations. SEM represents an advanced statistical technique beyond typical regression analysis. Regression analysis is a special type of SEM that typically focuses on understanding the relationship between one dependent observable variable and at least one independent observable variable. Unlike regression models, SEM allows multiple dependent variables to be included in the modelling simultaneously.[12–17] This enables researchers to explore not only direct relationships but also indirect relationships, often referred to as paths, among various observable or latent variables.[12–17] Therefore, SEM is particularly well suited for testing multiple hypotheses for various associations within a complex phenomenon. For example, several studies have adopted health behavioural theories, such as the health belief model (HBM), theory of planned behaviour (TPB), and extended parallel process model (EPPM), to investigate factors that encourage or discourage COVID-19 vaccine uptake.[12–17] In general, respondents are more likely to be vaccinated as a result of a perceived higher risk of being infected with the COVID-19 virus, perceive greater benefits of vaccines, or subjective norms.[12–17] Online surveys have been primarily used in SEM

research given their advantages such as cost effectiveness, easy administration, global outreach, and efficiency.[18] However, survey research involves the limitations of nonresponse bias, recall bias, assumed honesty, respondents' misunderstanding or misinterpretation of questions, and others.[18] Although social media data have been used in numerous COVID-19 social listening studies,[7,10] it is rare to find SEM studies that use social media data. Although researchers can apply ML or AI techniques to analyse large amounts of social media data, such studies do not demonstrate statistical relationships in the same way as SEM.

Accordingly, a new conceptual framework, social media infodemic listening (SoMeIL) for public health behaviour, has been proposed to address multifaceted health infodemics on social media.[19] The SoMeIL conceptual framework theorizes that social media users' online reaction behaviours can indicate their intentions to receive COVID-19 vaccines, for example.[19] In other words, parameters derived from social media platforms, such as the number of likes and shares of a given post, can be used as proxies for social media users' self-reported intentions towards COVID-19 vaccination in real life. Given our interest in social media and its critical role in health infodemics and thus people's behaviours, it is important to directly use social media data to validate such associations. SEM has been commonly used to validate conceptual frameworks where latent variables involve survey data,[12–17] but social media data have not been directly and extensively used in SEM analysis. Although many studies have investigated how social media has influenced people's intentions towards COVID-19 vaccination, most previous studies have relied on questionnaires to collect data,[12–17] while few studies have requested that participants provide their social media posts. The use of social media data is conceptually similar to typical SEM research with online surveys since social media data share the same benefits while mitigating some limitations. Ideally, researchers can retrieve as many relevant parameters and data as possible from application programming interfaces (APIs) on social media platforms. Thus, the sample size of social media data is generally not an issue. Social media data may have similar nonresponse biases due to inactive users or users not on a given social media platform, but this nonresponse bias can be addressed by using the numbers of likes, shares, or other parameters to infer the opinions of inactive social media users. In addition, when data from multiple social media platforms are collected for studies, it is possible to obtain a more comprehensive representation of the target audience. Since researchers do not need to design the questions, there is no need to assume respondents' honesty or worry about respondents misunderstanding or misinterpreting the questions. However, researchers need to actively screen posts as relevant or irrelevant after retrieving social media posts. Therefore, this study aims to validate partial components of the SoMeIL conceptual framework using SEM with Twitter data and demonstrates the feasibility of using Twitter data in SEM research.

Methods

Conceptual framework and hypotheses

The objective of this study was to preliminarily validate online reaction behaviours, intentions, and self-reported offline reaction behaviours in the proposed SoMeIL conceptual framework using SEM with Twitter data. Fig. 1 shows the proposed SEM derived from part of the SoMeIL conceptual framework and corresponding hypotheses. Directly measured variables are represented by rectangles, and latent variables are represented by circles. The definitions of the key terms shown in Fig. 1 are presented below.

- Sentiment_score: a continuous value normalized between −1 (most negative) and +1 (most positive) by summing positive, negative and neutral scores via the Valence Aware Dictionary and sEntiment Reasoner (VADER) for each tweet.[20]
- Favourite_log: transformed into favourite_log from favourite_count, which represents the number of times that a tweet was liked by Twitter users.[21]
- Retweet_log: transformed into retweet_log from retweet_count, which represents the number of times a tweet has been retweeted (i.e., shared).[21]
- Tweet engagement: a latent variable that represents engagement activities inferred at the tweet level.
- User_favourites_log: transformed into user_favourites_log from user_favourites_count, which represents the number of followers the account currently has.[22]
- User_followers_log: transformed into user_followers_log from user_followers_count, which represents the number of followers the account currently has.[22]
- User_friends_log: transformed into user_friends_log from user_friends_count, which is the number of users the account is following (i.e., "followings").[22]
- User_listed_log: transformed into user_listed_log from user_listed_count, which represents the number of public lists that the user is a member of.[22] It is transformed into user_listed_log.
- User engagement: a latent variable that represents engagement activities inferred at the user level.
- Vaccinated: a tweet that indicates a Canadian Twitter user's intention to receive the first dose of the COVID-19 vaccine.

The health information in this case is from the massive vaccination campaign that encouraged people in Canada to receive the first dose of the COVID-19

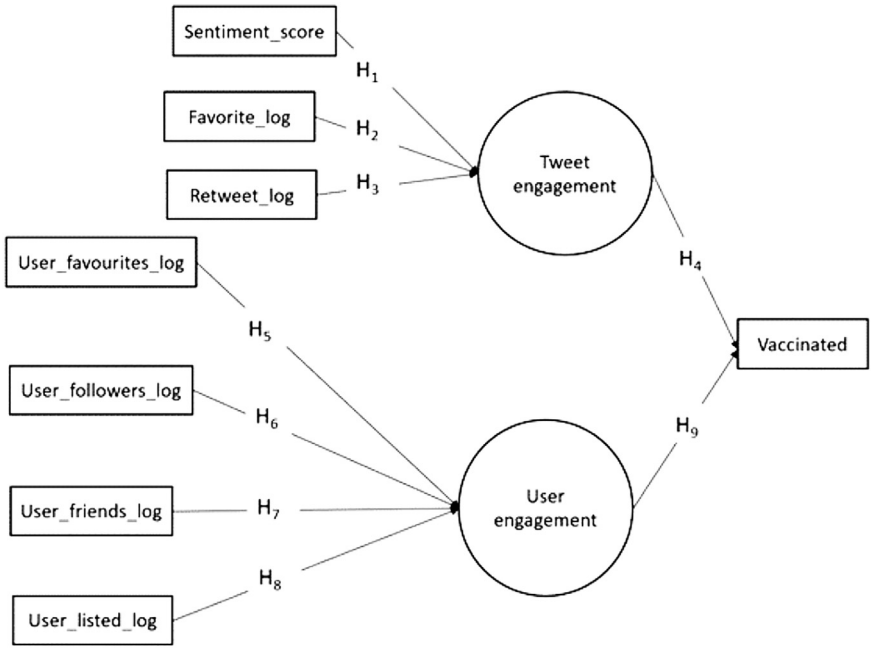

Fig. 1: Proposed social media infodemic listening for public health behaviour conceptual framework using Twitter data. **The components are: Sentiment score (a continuous value normalized between -1 as most negative and +1 as most positive), Favorite_log (natural logarithm transformation of favorite counts), Retweet_log (natural logarithm transformation of retweet counts), User_favourite_log (natural logarithm transformation of a user's favourite counts), User_followers_log (natural logarithm transformation of a user's follower counts), User_friends_log (natural logarithm transformation of a user's friend counts), User_listed_log (natural logarithm transformation of the number of public lists that the user is a member of), User engagement (a latent variable that represents engagement activities inferred at the user level), Vaccinated (a tweet indicating a user's intention to receive the first dose of the COVID-19 vaccine), H_1 (Hypothesis 1), H_2 (Hypothesis 2), H_3 (Hypothesis 3), H_4 (Hypothesis 4), H_5 (Hypothesis 5), H_6 (Hypothesis 6), H_7 (Hypothesis 7), H_8 (Hypothesis 8), and H_9 (Hypothesis 9).**

vaccine. Online reaction behaviours include sentiment scores (i.e., emotion in the framework), the log-number of favourites, the log-number of retweets, the log-number of user favourites, the log-number of user followers, the log-number of user friends, and the log-number of times a user is listed in a tweet. Offline reaction behaviour is self-reported vaccination or not in a tweet. We theorized positive associations for all hypotheses as follows:

- H_1: There is a significant relationship between a tweet's sentiment score and tweet engagement.
- H_2: There is a significant relationship between the log-number of a tweet's favourites and tweet engagement.
- H_3: There is a significant relationship between the log-number of a tweet's retweets and tweet engagement.
- H_4: There is a significant relationship between tweet engagement and COVID-19 vaccination.
- H_5: There is a significant relationship between the log-number of a user's favourites and user engagement.
- H_6: There is a significant relationship between the log-number of a user's followers and user engagement.
- H_7: There is a significant relationship between the log-number of a user's friends and user engagement.
- H_8: There is a significant relationship between the log-number of a user's public lists and user engagement.
- H_9: There is a significant relationship between user engagement and COVID-19 vaccination.

Data collection

This study utilized a cross-sectional design since we were interested in understanding COVID-19 vaccination behaviours among adults in Toronto and Ottawa when the first dose of COVID-19 vaccines became available via online appointments. English tweets related to the COVID-19 pandemic from March 8 to June 30, 2021, were retrieved via Twitter's Academic API using the keywords and hashtags listed in Table S1 (Supplementary Materials). This process resulted in approximately two billion tweets. Next, the tweets were narrowed to those that included "Toronto" or "Ottawa"

in tweets or in users' locations to gather as many tweets as possible in Toronto or Ottawa, Ontario, Canada. This approach was used to address missing geolocations indicated in the literature[23] and resulted in approximately four million tweets.

To prepare for the subsequent sentiment analysis, Twitter handles (i.e., @username), uniform resource locator links (URLs), punctuation, stop words, and retweets were removed in accordance with existing studies.[24,25] Then, the words in a tweet were converted to their most general form[24,25] using the Natural Language Toolkit (NLTK) package version 3.8.1.[26]

Measures

In addition to tweets, the other directly measured independent variables were added and then transformed using the natural logarithm given the presence of zeros since the natural logarithm of one is zero. The variables were subsequently grouped to represent the latent dependent variables, tweet engagement and user engagement, as shown in Fig. 1 and based on the SoMeIL conceptual framework.

To prepare for the dependent variable "vaccinated" shown in Fig. 1, a subset of the four million tweets was created by retrieving tweets that included "appoint," "jab," "shot," and "vaccin." We manually reviewed and labelled tweets "1" if users explicitly self-reported that they were seeking or waiting for a vaccine appointment or if they were already vaccinated with the COVID-19 vaccine. Tweets were labelled "0" if users explicitly self-reported that they were hesitant or against the COVID-19 vaccine. Other tweets were excluded if they did not include explicit expressions about the COVID-19 vaccination or if they were news, although these were still relevant to the overall pandemic and vaccine rollout in Canada. The subset ultimately included 2420 English tweets with 2420 unique users, which was comparable to the sample sizes of survey respondents in the existing SEM literature.[12–17]

Statistical analysis

Descriptive statistics, such as means or frequencies, standard deviations, and Spearman correlations, were used to describe the measures in the proposed model (Supplementary Materials Tables S2 and S3, respectively), except for the latent variables. Spearman correlations were calculated to account for outliers and nonnormal distributions in some measured variables even after the data transformation via the natural logarithm (Supplementary Materials Appendix A).

Confirmatory factor analysis (CFA) with diagonally weighted least squares (DWLS), also known as robust WLS, was used to test the "fit" of the observed variables for each latent variable. The robust WLS was specified because some measured variables still violated the normal distribution assumption after the data transformation.[27–29] For each CFA model, variables were removed until the fit indices, including chi-square, comparative fit index (CFI), goodness of fit (GFI), adjusted goodness of fit (AGFI), Tucker–Lewis index (TLI), and root mean square error of approximation (RMSEA), were acceptable. For the CFI, GFI, AGFI and TLI, $\geq$0.90 is generally considered acceptable and $\geq$0.95 is considered good. An RMSEA $\leq$0.08 is recommended.[30,31]

After CFA, SEM was performed to test the proposed model (Model 1) in Fig. 1 with the DWLS and the same recommended criteria for the fit indices. Model 1 was optimized if the model fit indices suggested that better models could be found according to the proposed conceptual framework and correlation matrix. All the data analyses were performed in Juypter Notebook, available in Anaconda version 4.3.3, with the semopy package used for CFA and SEM.[32,33]

Ethical approval

This study was approved by the University of Waterloo Office of Research Ethics (#43961).

Role of funding source

This study was supported by the 2023-24 Ontario Graduate Scholarship awarded by the Government of Ontario in Canada. The funding source had no role in study design, data collection/analyses/interpretation, manuscript preparation, or submission at all. **All authors had full access to all of the study data and took final responsibility for the decision to submit for publication.**

Results

The descriptive statistics and correlations of the measured variables are shown in Table S2 and Table S3 (Supplementary Materials), respectively. The results of CFA are shown in Table 1. The latent variable "tweet_engagement" was saturated, and the latent variable "user_engagement" had good fit indices except for the RMSEA, which was greater than the recommended 0.08. When both latent variables were combined in the full measurement model, CFA revealed borderline fit indices that were close to the acceptable cut-off points. The RMSEA of the full-measure model also decreased slightly.

Given the borderline CFA results using DWLS and Twitter data instead of typical surveys, we decided to test Model 1 using SEM. Fig. 2 presents Model 1, and the model fit indices are shown in Table 1. As Fig. 2 illustrates, two hypotheses, H_2 and H_6, were not supported because they did not have a statistically significant association. Instead, SEM suggested that the log-number of a tweet's favourites and the log-number of a user's followers were fixed in the model as references:

- H_1: There is a significant relationship between a tweet's sentiment score and tweet engagement ($p < 0.05$).

	Tweet_engagement	User_engagement	Complete measurement model
Degrees of freedom	0	2	13
chi-square p value[a]	–	<0.05	<0.05
CFI	0.98	0.97	0.88
GFI	0.98	0.97	0.88
AGFI	–	0.90	0.80
NFI	0.98	0.97	0.88
TLI	–	0.91	0.80
RMSEA	∞	0.12	0.12

CFI: comparative fit index. GFI: goodness of fit. AGFI: adjusted goodness of fit. NFI: Non-Normed Fit Index. TLI: Tucker–Lewis index. RMSEA: root mean square error of approximation. [a]The chi-squared p value is not recommended for consideration regardless of the SEM because it is heavily influenced by the sample size.

Table 1: **Fit statistics for each latent variable and full measurement model.**

- H_2: There is a significant relationship between the log-number of a tweet's favourites and tweet engagement (p value not provided).
- H_3: There is a significant relationship between the log-number of a tweet's retweets and tweet engagement ($p < 0.05$).
- H_4: There is a significant relationship between tweet engagement and COVID-19 vaccination ($p < 0.05$).
- H_5: There is a significant relationship between the log-number of a user's favourites and user engagement ($p < 0.05$).
- H_6: There is a significant relationship between the log-number of a user's followers and user engagement (p-value not provided).
- H_7: There is a significant relationship between the log-number of a user's friends and user engagement ($p < 0.05$).
- H_8: There is a significant relationship between the log-number of a user's public lists and user engagement ($p < 0.05$).
- H_9: There is a significant relationship between user engagement and COVID-19 vaccination ($p < 0.05$).

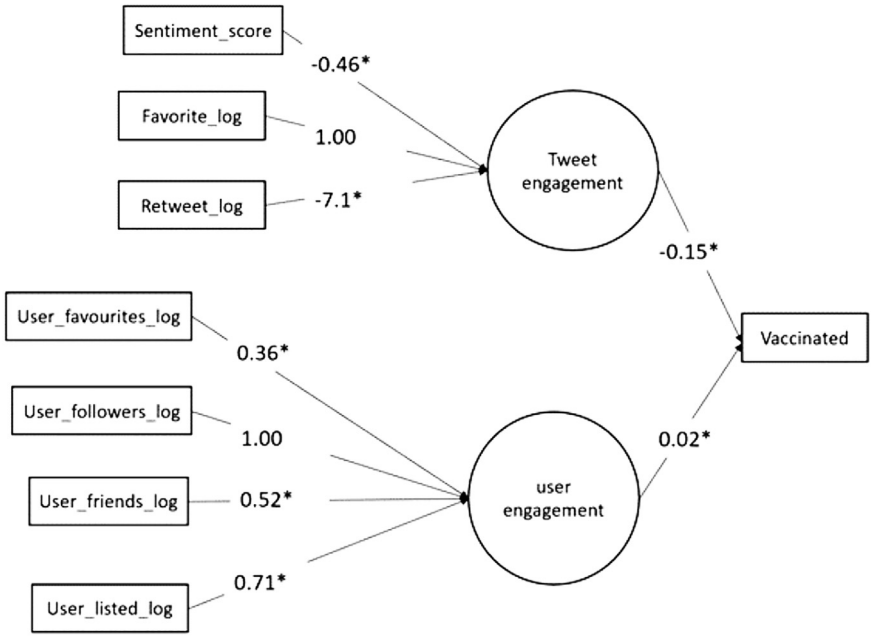

Fig. 2: Model 1 results according to the proposed conceptual framework. ***p < 0.05. The components are: Sentiment score (a continuous value normalized between -1 as most negative and +1 as most positive), Favorite_log (natural logarithm transformation of favorite counts), Retweet_log (natural logarithm transformation of retweet counts), User_favourite_log (natural logarithm transformation of a user's favourite counts), User_followers_log (natural logarithm transformation of a user's follower counts), User_friends_log (natural logarithm transformation of a user's friend counts), User_listed_log (natural logarithm transformation of the number of public lists that the user is a member of), User engagement (a latent variable that represents engagement activities inferred at the user level), Vaccinated (a tweet indicating a user's intention to receive the first dose of the COVID-19 vaccine).**

The fit indices of Model 1 in Table 2 indicated that the model could be optimized. According to the results from CFA and SEM for Model 1, it was hypothesized that instead of two latent variables, one might be better. Fig. 3 shows the final SEM (Model 2) after model revisions based on the proposed conceptual framework. That is, instead of two latent variables representing engagement activities at the tweet and user levels, one latent variable, "VaxIntent," was proposed to represent Canadian Twitter users' intentions to be vaccinated against COVID-19. The model indices of Model 2 are also included in Table 2. According to Model 2, the log-number of a user's friends was removed, and the remaining variables had statistically significant relationships with the latent variable. Nonetheless, it was not straightforward to interpret the estimated coefficients and standard errors when the variables were transformed with the natural logarithm. Therefore, Tables S4 and S5 (Supplementary Materials) show the coefficients and standard errors for each variable in Model 1 and Model 2, respectively, after the estimates were converted back.

Discussion

The present study was conducted to preliminarily evaluate the online reaction behaviours, emotions, intentions, and self-reported offline behaviours proposed in the SoMeIL conceptual framework.[19] According to the SoMeIL conceptual framework,[19] sentiment scores as emotions, the log numbers of a tweet's favourites and retweets, and users' favourites, followers, and public lists, as online reaction behaviours, were investigated using SEM to assess their relationships with self-reported COVID-19 vaccination, as an offline reaction behaviour, with a total of 2420 English tweets. As shown in Table S4 (Supplementary Materials), most variables in Model 1 had positive associations. However, the relationships between a tweet's sentiment score and tweet engagement, between the number of a tweet's retweets and tweet engagement, and between tweet engagement and vaccination could vary. Similarly, in Model 2, the association between the number of a user's followers and COVID-19 vaccination intention could be positive or negative (Table S5 in the Supplementary Materials), whereas other variables in Model 2 had positive associations with the latent variable. However, Model 2 was the best model according to the fit indices shown in Table 2, and all the variables in Model 2 had statistically significant relationships despite one unstable variable. According to Model 2 (Table S5 in the Supplementary Materials), the sentiment score had the strongest positive relationship with COVID-19 vaccination intention, followed by the number of public lists to which a user belonged. The number of followers had the weakest association with COVID-19 vaccination intention.

	Model 1	Model 2
Degrees of freedom	18	20
Chi-square p value[a]	P < 0.05	1.0000
CFI	0.8321	1.0079
GFI	0.8287	1.0000
AGFI	0.7335	1.0000
NFI	0.8267	1.0000
TLI	0.7388	1.0106
RMSEA	0.1219	0.0000

CFI: comparative fit index. GFI: goodness of fit. AGFI: adjusted goodness of fit. NFI: Non-Normed Fit Index. TLI: Tucker–Lewis index. RMSEA: root mean square error of approximation. [a]The chi-squared p value is not recommended for consideration regardless of the SEM because it is heavily influenced by sample size.

Table 2: **Model fit indices for Model 1 and Model 2.**

Overall, Model 2 provides preliminary results that validate the partial components of the SoMeIL conceptual framework given the significant associations. That is, variables derived from Twitter could be used to infer Twitter users' intentions to receive the COVID-19 vaccine, which was the latent variable. The sentiment score, which was calculated via VADER sentiment analysis,[23] represented emotions and was significantly associated with the literature.[2,12–17] In other words, Twitter users who generally expressed positive sentiments towards the COVID-19 vaccine were more likely to be vaccinated again during the pandemic.[2,12–17] The other variables exhibited similar relationships. The more favourites and retweets a tweet received or the more favourites, followers, or public lists a user received, the more likely the user was to accept the first dose of the COVID-19 vaccine, although the number of a user's followers could have a negative effect in some cases.

Surprisingly, it appeared that outliers had little impact on SEM since Model 2 met all the recommended criteria of the fit indices. In fact, when outliers were removed or replaced with medians, none of the structural equation models converged. This outcome remained unchanged even after different combinations of the measured variables were tested. For example, "favourite_log" was excluded because it became useless after its outliers were removed or replacing with its median, which was zero. This approach allowed the variable to include only zeros since nonzero values were outliers. Even after "favourite_log" was excluded, the other SEMs still failed to converge. Therefore, we hypothesized that without the "favourite_log" variable, the remaining data would not fit the SEM well.[34,35] Therefore, although the assumption of no outliers in SEM was violated in the current study, the outliers actually included important information that should not be removed from the modelling. Given the nature of social media data, outliers could be legitimate since some tweets could receive more likes or shares or some users could have more followers or likes than others.

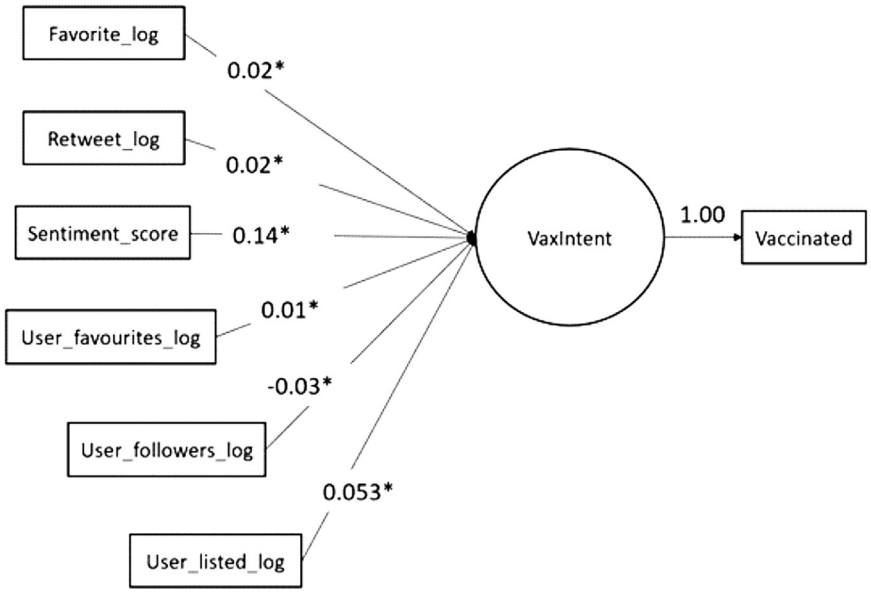

Fig. 3: Model 2 results after Model 1 is optimized. ***p < 0.05. The components are: Sentiment score (a continuous value normalized between -1 as most negative and +1 as most positive), Favorite_log (natural logarithm transformation of favorite counts), Retweet_log (natural logarithm transformation of retweet counts), User_favourite_log (natural logarithm transformation of a user's favourite counts), User_followers_log (natural logarithm transformation of a user's follower counts), User_friends_log (natural logarithm transformation of a user's friend counts), User_listed_log (natural logarithm transformation of the number of public lists that the user is a member of), User engagement (a latent variable that represents engagement activities inferred at the user level), Vaccinated (a tweet indicating a user's intention to receive the first dose of the COVID-19 vaccine).**

In addition to the preliminary validation of the partial components within the SoMeIL conceptual framework, this study may be the first to use only Twitter data in SEM research. The findings show promise for the use of Twitter data in SEM research with proper theoretical frameworks, but there are several limitations. First, the generalizability of this study was limited since it did not include Twitter users who were excluded from the data or non-Twitter users. Furthermore, SEM was conducted in a cross-sectional manner, so it offered only a snapshot of the entire pandemic. In the future, longitudinal SEM could be performed. However, unlike surveys, researchers have no control over the frequency of people's tweeting behaviours. Some very active users might tweet daily, whereas others might tweet sporadically. Considerable effort would be required to find enough users with similar tweeting frequencies to conduct a longitudinal SEM study, although this would not be impossible. The quality of the data was another major limitation. For example, users' demographic information, such as sex and gender, was not available to the researchers unless users self-identified their demographic information on their Twitter profiles. Extensive manual identification or complex ML or AI techniques are required to retrieve or infer users' complete demographic characteristics from Twitter data.[36,37] This could lead to even fewer representative samples since the majority of Twitter users do not include demographic information in their profiles. Additionally, there are other methods for calculating sentiments,[8,9] although VADER sentiment analysis has commonly been used.[23] The data transformation via the natural logarithm also limited the data quality due to information loss. In general, log transformations are not recommended for count data despite their common usage in linear models such as regressions and SEM.[34,35] Instead, modelling count data with Poisson or negative binomial distributions is recommended.[35] Nonetheless, Poisson or negative binomial distributions have not been made available in open-source SEM packages, such as the semopy package.[32,33] We mitigated this concern by using the DWLS to analyse data that did not meet the normal distribution assumption.[27–29,34,35] Finally, ecological fallacy is a disadvantage in a SEM study. In other words, the findings should not be interpreted at the individual level.

Despite these limitations, this study confirmed that Twitter data can be useful for SEM research and partially and preliminarily validated the SoMeIL conceptual framework.[19] That is, parameters retrieved from Twitter as online reaction behaviours can be used to infer Twitter users' self-reported intentions, which can be used as a proxy for users' vaccination behaviours in real life. For future research, we plan to apply ML or AI techniques to correctly classify self-reported offline

reaction behaviours to scale up the data sample. Alternatively, instead of self-reported offline reaction behaviours derived from Twitter, other data on offline reaction behaviours can be collected and analysed to further validate the SoMeIL framework. Alternatively, different social media data can be collected, such as videos and images, to study how the SEM approach and the SoMeIL conceptual framework can be applied. For example, as in typical SEM research, future studies can design questionnaires to collect participants' demographic information and request that participants voluntarily give social media posts to researchers to investigate how participants' online and offline reaction behaviours are associated with their demographic information. However, we acknowledge that collecting social media has become increasingly difficult for researchers since social media platforms have started to restrict their API access.

This study provided preliminary validations of parts of the SoMeIL conceptual framework. The results showed that the six variables retrieved from Twitter had statistically significant relationships with the latent variable, which could be used as a proxy for Twitter users' self-reported COVID-19 vaccination uptake. This study also demonstrated that it is feasible to use Twitter data in SEM research. However, further studies are needed to examine other SEM approaches and other social media platforms to further validate the SoMeIL conceptual framework. As social media have been integrated into people's daily lives worldwide, their dominance will increase the impact of health infodemics. As a result, in addition to conventional channels such as surveys or word of mouth, it is crucial to "listen to" public discourse on different social media platforms and address emerging confusion, questions, and even misinformation in a timely manner.[10] As this study illustrates, several indicators on Twitter, such as the numbers of likes and shares, can be used to infer the vaccination behaviours of Toronto and Ottawa Twitter users in real life. Therefore, this approach can be adopted to forecast vaccination coverage for future vaccine-preventable diseases. This approach can also help tailor communication strategies and address specific issues based on Twitter users' discussions and online behaviours to effectively reach different groups.[38,39] The SoMeIL conceptual framework can be extended to other areas, such as symptom reports or behavioural patterns, to aid in public health decision-making and resource allocations.[40] By integrating social media platforms such as Twitter into pandemic preparedness, health organizations and government authorities can harness their potential as powerful tools to engage with the public, address health misinformation, and effectively respond to crises, which can ultimately help to mitigate the impact of future pandemics.[38,39] Similar to the WHO's EARS platform,[10] the SoMeIL conceptual framework can be implemented as a way to provide real-time monitoring and surveillance. The literature has shown that social media can be used for the early detection of emerging health threats and to track misinformation trends.[7,10,11] Social media data can also complement traditional surveillance methods and help public health authorities respond quickly to potential outbreaks.

Overall, this study provides a preliminary yet quantifiable method to examine social listening based on components of the SoMeIL conceptual framework. It is recommended that future pandemic preparedness recognize the substantial roles of social media in shaping public perception, disseminating information, and influencing behaviours during a health crisis. Incorporating social media into pandemic preparedness strategies can enhance communication, information sharing, and response efforts.

Contributors

Shu-Feng Tsao has conceptualised the study, collected and analysed data. She has also drafted and revised the manuscript. Dr. Chen and Dr. Butt have supervised the methodology of the study and reviewed the manuscript draft. All authors reviewed the results and approved the final version of the manuscript. All authors have accessed and verified the data used in the study.

Data sharing statement

The subset and codes used in this study are available on GitHub.

Declaration of interests

The authors have no conflicts of interest although the first author has received the 2023-24 Ontario Graduate Scholarship for this study.

Acknowledgements

This study is supported by the 2023-24 Ontario Graduate Scholarship awarded to the first author by the Government of Ontario, Canada. The funding source has no involvement in the study at all.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102544.

References

1 Cascini F, Pantovic A, Al-Ajlouni YA, et al. Social media and attitudes towards a COVID-19 vaccination: a systematic review of the literature. *eClinicalMedicine*. 2022;48(101454):101454. https://doi.org/10.1016/j.eclinm.2022.101454.
2 Romate J, Rajkumar E, Gopi A, et al. What contributes to COVID-19 vaccine hesitancy? A systematic review of the psychological factors associated with COVID-19 vaccine hesitancy [cited 2023 Jul 22] *Vaccines (Basel)*. 2022;10(11):1777. Available from: https://www.mdpi.com/2076-393X/10/11/1777.
3 Skafle I, Nordahl-Hansen A, Quintana DS, Wynn R, Gabarron E. Misinformation about COVID-19 vaccines on social media: rapid review. *J Med Internet Res*. 2022;24(8):e37367. https://doi.org/10.2196/37367.
4 Zhao S, Hu S, Zhou X, et al. The prevalence, features, influencing factors, and solutions for COVID-19 vaccine misinformation: systematic review. *JMIR Public Health Surveill*. 2023;9:e40201. https://doi.org/10.2196/40201.
5 Thorakkattil SA, Abdulsalim S, Karattuthodi MS, Unnikrishnan MK, Rashid M, Thunga G. COVID-19 vaccine hesitancy: the perils of peddling science by social media and the lay press. *Vaccines (Basel)*. 2022;10(7):1059. https://doi.org/10.3390/vaccines10071059.
6 Lieneck C, Heinemann K, Patel J, et al. Facilitators and barriers of COVID-19 vaccine promotion on social media in the United States: a systematic review [cited 2023 Jul 22] *Healthcare (Basel)*.

2022;10(2):321. Available from: https://www.mdpi.com/2227-9032/10/2/321.
7 Butt MJ, Malik AK, Qamar N, Yar S, Malik AJ, Rauf U. A survey on COVID-19 data analysis using AI, IoT, and social media. *Sensors*. 2023;23(12):5543. https://doi.org/10.3390/s23125543.
8 Alamoodi AH, Zaidan BB, Al-Masawa M, et al. Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy. *Comput Biol Med*. 2021;139(104957):104957. https://doi.org/10.1016/j.compbiomed.2021.104957.
9 Umair A, Masciari E, Habib Ullah MH. Sentimental analysis applications and approaches during COVID-19: a survey. In: *25th international database engineering & applications symposium*. New York, NY, USA: ACM; 2021. https://doi.org/10.1145/3472163.3472274.
10 Purnat TD, Wilson H, Nguyen T, Briand S. Ears – a WHO platform for AI-supported real-time online Social Listening of COVID-19 conversations. In: *Studies in health technology and informatics*. IOS Press; 2021. https://doi.org/10.3233/SHTI210330.
11 Heyerdahl LW, Lana B, Giles-Vernick T. Rethinking the infodemic: social media and offline action in the COVID-19 pandemic. In: *Economics, law, and institutions in Asia Pacific*. Singapore: Springer Singapore; 2022:73–82.
12 Chu H, Liu S. Integrating health behavior theories to predict American's intention to receive a COVID-19 vaccine. *Patient Educ Couns*. 2021;104(8):1878–1886. https://doi.org/10.1016/j.pec.2021.02.031.
13 Fan C-W, Chen I-H, Ko N-Y, et al. Extended theory of planned behavior in explaining the intention to COVID-19 vaccination uptake among mainland Chinese university students: an online survey study. *Hum Vaccin Immunother*. 2021;17(10):3413–3420. https://doi.org/10.1080/21645515.2021.1933687.
14 Irfan M, Shahid AL, Ahmad M, et al. Assessment of public intention to get vaccination against COVID -19: evidence from a developing country. *J Eval Clin Pract*. 2022;28(1):63–73. https://doi.org/10.1111/jep.13611.
15 Mir HH, Parveen S, Mullick NH, Nabi S. Using structural equation modeling to predict Indian people's attitudes and intentions towards COVID-19 vaccination. *Diabetes Metab Syndr*. 2021;15(3):1017–1022. https://doi.org/10.1016/j.dsx.2021.05.006.
16 Bui HN, Duong CD, Nguyen VQ, et al. Utilizing the theory of planned behavior to predict COVID-19 vaccination intention: a structural equational modeling approach. *Heliyon*. 2023;9(6):e17418. https://doi.org/10.1016/j.heliyon.2023.e17418.
17 Drążkowski D, Trepanowski R. Reactance and perceived disease severity as determinants of COVID-19 vaccination intention: an application of the theory of planned behavior. *Psychol Health Med*. 2022;27(10):2171. https://doi.org/10.1080/13548506.2021.2014060.
18 Evans JR, Mathur A. The value of online surveys: a look back and a look ahead. *Internet Res*. 2018;28(4):854–887. https://doi.org/10.1108/intr-03-2018-0089.
19 Tsao S-F. *SoMeIL: a social media infodemic listening for public health behaviours conceptual framework*. University of Waterloo; 2023. Available from: http://hdl.handle.net/10012/20029.
20 Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. *Proc Int AAAI Conf Weblogs Soc Media*. 2014;8(1):216–225. https://doi.org/10.1609/icwsm.v8i1.14550.
21 Tweet object. Twitter [cited 2023 Jul 23]. Available from: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet.
22 User object. Twitter [cited 2023 Jul 23]. Available from: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/user.
23 Huang B, Carley KM. A large-scale empirical study of geotagging behavior on Twitter. In: *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*. New York, NY, USA: ACM; 2019. https://doi.org/10.1145/3341161.3342870.
24 Liu S, Liu J. Public attitudes toward COVID-19 vaccines on English-language Twitter: a sentiment analysis. *Vaccine*. 2021;39(39):5499–5505. https://doi.org/10.1016/j.vaccine.2021.08.058.
25 Reshi AA, Rustam F, Aljedaani W, et al. COVID-19 vaccination-related sentiments analysis: a case study using worldwide Twitter dataset. *Healthcare (Basel)*. 2022;10(3):411. https://doi.org/10.3390/healthcare10030411.
26 Bird S, Klein E, Loper E. *Natural language processing with python: analyzing text with the natural language toolkit*. O'Reilly Media; 2009.
27 Whittaker TA, Schumacker RE. *A beginner's guide to structural equation modeling*. 5th ed. London, England: Routledge; 2022.
28 Bowen NK, Guo S. Evaluating and improving CFA and general structural models. In: *Structural equation modeling*. Oxford University Press; 2011:135–166.
29 Kupek E. Beyond logistic regression: structural equations modelling for binary variables and its application to investigating unobserved confounders. *BMC Med Res Methodol*. 2006;6(1). https://doi.org/10.1186/1471-2288-6-13.
30 Hooper D, Coughlan J, Mullen MR. Structural equation modelling: guidelines for determining model fit [cited 2023 Jul 29] *Electron J Bus Res Methods*. 2008;6(1):53–60. Available from: https://academic-publishing.org/index.php/ejbrm/article/view/1224.
31 Peugh J, Feldon DF. "How well does your structural equation model fit your data?": is marcoulides and Yuan's equivalence test the answer? *CBE Life Sci Educ*. 2020;19(3):es5. https://doi.org/10.1187/cbe.20-01-0016.
32 Igolkina AA, Meshcheryakov G. Semopy: a python package for structural equation modeling. *Struct Equ Model*. 2020;27(6):952–963. https://doi.org/10.1080/10705511.2019.1704289.
33 Meshcheryakov G, Igolkina AA, Samsonova MG. *Semopy 2: a structural Equation Modeling package with random effects in Python*. 2021. https://doi.org/10.48550/ARXIV.2106.01140.
34 Xia Y, Yang Y. RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: the story they tell depends on the estimation methods. *Behav Res Methods*. 2019;51(1):409–428. https://doi.org/10.3758/s13428-018-1055-2.
35 Green JA. Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression. *Health Psychol Behav Med*. 2021;9(1):436–455. https://doi.org/10.1080/21642850.2021.1920416.
36 Golder S, Stevens R, O'Connor K, James R, Gonzalez-Hernandez G. Methods to establish race or ethnicity of Twitter users: scoping review. *J Med Internet Res*. 2022;24(4):e35788. https://doi.org/10.2196/35788.
37 Cesare N, Grant C, Nsoesie EO. Detection of user demographics on social media: a review of methods and recommendations for best practices. preprint arXiv:1702.01807 *arXiv*. 2017:1–25.
38 Berg SH, O'Hara JK, Shortt MT, et al. Health authorities' health risk communication with the public during pandemics: a rapid scoping review. *BMC Public Health*. 2021;21(1). https://doi.org/10.1186/s12889-021-11468-3.
39 Vraga EK, Jacobsen KH. Strategies for effective health communication during the Coronavirus pandemic and future emerging infectious disease events. *World Med Health Policy*. 2020;12(3):233–241. https://doi.org/10.1002/wmh3.359.
40 Yang Y, Tsao S-F, Basri MA, Chen HH, Butt ZA. Digital disease surveillance for emerging Infectious Diseases: an early warning system using the internet and social media data for COVID-19 forecasting in Canada. In: *Caring is sharing – exploiting the value in data for health and innovation*. IOS Press; 2023. https://doi.org/10.3233/shti230290.