A critique of "Polarization of Opinions on COVID-19 Measures - Integrating Twitter and Survey Data"

Markus Reiter-Haas, Beate Klosch, Markus Hadler, and Elisabeth Lex

September 19th, 2022

Presented by Jeff Brozena

Objectives

1. Identify main concepts and research questions
2. Description of methods
3. Interpretation of results
4. Critique of methods
5. Discussion

Summary

The authors analyze Twitter data, surveys, and an integrated combination of the two in order to investigate the level of opinion polarization surrounding COVID-19 prevention measures in the German-speaking DACH region.

Survey data and Twitter data are temporally aligned, and a subset of Twitter data is authored by survey respondents. An interesting effort is made to avoid the ecological fallacy, which occurs when analyzing correlations in aggregate data alongside correlations in the data of individuals.

As a proxy for opinion affect, the authors performed sentiment analysis on Twitter data, estimating opinions on a scale ranging from -1 to 1, with -1 representing negative affect. Human annotators then assign an agreement score to each Twitter account's tweets to evaluate congruence of opinions on Twitter and survey answers.

Summary

The authors analyze Twitter data, surveys, and an integrated combination of the two in order to investigate the level of opinion polarization surrounding COVID-19 prevention measures in the German-speaking DACH region.

Survey data and Twitter data are temporally aligned, and a subset of Twitter data is authored by survey respondents. An interesting effort is made to avoid the ecological fallacy, which occurs when analyzing correlations in aggregate data alongside correlations in the data of individuals.

As a proxy for opinion affect, the authors performed sentiment analysis on Twitter data, estimating opinions on a scale ranging from -1 to 1, with -1 representing negative affect. Human annotators then assign an agreement score to each Twitter account's tweets to evaluate congruence of opinions on Twitter and survey answers.

Summary

The authors analyze Twitter data, surveys, and an integrated combination of the two in order to investigate the level of opinion polarization surrounding COVID-19 prevention measures in the German-speaking DACH region.

Survey data and Twitter data are temporally aligned, and a subset of Twitter data is authored by survey respondents. An interesting effort is made to avoid the ecological fallacy, which occurs when analyzing correlations in aggregate data alongside correlations in the data of individuals.

As a proxy for opinion affect, the authors performed sentiment analysis on Twitter data, estimating opinions on a scale ranging from -1 to 1, with -1 representing negative affect. Human annotators then assign an agreement score to each Twitter account's tweets to evaluate congruence of opinions on Twitter and survey answers.

Summary

So what?

The authors note that high levels of public opinion polarization, especially around political party lines, can promote adverse effects like hostility.

Concept and Research Question

Opinion Polarization

Opinion Polarization

Empirically defined as a **state** where public opinions are characterized by extreme positions

Opinion Polarization

Empirically defined as a **state** where public opinions are characterized by extreme positions

Operationalized as **extracted sentiment** in Twitter data and as **expressed agreement with opinions on polarizing topics** in survey data

Opinion Polarization

Empirically defined as a **state** where public opinions are characterized by extreme positions

Operationalized as **extracted sentiment** in Twitter data and as **expressed agreement with opinions on polarizing topics** in survey data

Measured using **dispersion** and **modality** of opinions

Opinion Polarization

Empirically defined as a **state** where public opinions are characterized by extreme positions

Operationalized as **extracted sentiment** in Twitter data and as **expressed agreement with opinions on polarizing topics** in survey data

Measured using **dispersion** and **modality** of opinions

Dispersion is *quantified* using **variance**, a value's distance from an average

Opinion Polarization

Empirically defined as a **state** where public opinions are characterized by extreme positions

Operationalized as **extracted sentiment** in Twitter data and as **expressed agreement with opinions on polarizing topics** in survey data

Measured using **dispersion** and **modality** of opinions

Dispersion is *quantified* using **variance**, a value's distance from an average

Modality is *quantified* using **kurtosis**, the sharpness of the peak of a frequency distribution curve

Research Question

For German-speaking members of DACH countries, what is the relationship between COVID-19 prevention measures and opinion polarization?

Independent variable: COVID-19 prevention measures

Dependent variable: Opinion polarization

Methods

Dataset Summary

The authors analyze opinion polarization in three sources:

1. Twitter data using an **open dataset** of tweet IDs
2. Survey responses collected from a representative **online survey**
3. An **integrated dataset** containing survey responses and historic tweets of those respondents who shared their Twitter handle

Dataset Summary

A total of six perspectives become available following a **subsetting** of each source:

1. A **Twitter** subset is made up of tweets with temporal overlap between the open dataset and the survey respondents' historic tweets.
2. A **survey** subset is made up of respondents who self-report actively using Twitter.
3. An **integrated** Twitter/survey subset is made up of survey respondents who provided their Twitter handle.

Levels of Measurement

- The **open dataset** arguably uses a ratio scale (contains a zero point and ranked, equal intervals)
- The **survey** uses an ordinal scale from 1 as strong disagreement to 5 as strong agreement
- The **integrated dataset** uses an ordinal rating scale of agreement, similar to the survey

Twitter Dataset

Tweets are retrieved from Twitter's streaming API and filtered through several passes.

Pass #	Filter	Tweet Count
First	1% sample of tweets from Twitter streaming API	Not reported
Second	Include only German language	3,336,562
Third	Include only tweets from survey time period	567,579

The third pass results in what the authors refer to as the "subset" of open Twitter data.

Twitter Dataset

A fourth filtering pass included tweets containing word stems: *impf* for vaccination, *mask* for mask wearing, and *trac* for contact tracing. The authors note that this step included "virtually all tweets related to these [preventative] measures" [1]

Word Stem	Full Tweet Count	Subset Tweet Count
impf	63,676	12,260
mask	136,198	31,856
trac	13,151	1,385

[1] Reiter-Haas, Klösch, Hadler, & Lex (2022)

Finally, sentiment analysis was performed using the `TextBlob` library with a German language extension containing a sentiment polarity lexicon. Sentiment is extracted on a scale of -1 to 1, where -1 is absolutely negative sentiment and 1 is absolutely positive.

After filtering out purely objective (i.e., scientific or factual) statements, **the final dataset is as follows**.

Prevention Measure	Full Tweet Count	Subset Tweet Count
Vaccination	25,769	5,420
Masking	60,218	15,425
Contact tracing	4,819	634

Survey Dataset

The authors collected survey responses concerning individuals' socio-demographics, social media behaviors, and opinions on COVID-19 prevention measures.

The survey was conducted from July 30th, 2020 and ultimately concluded on August 10th, 2020.

For context, the authors provide values the stringency index, a measure of strictness of active COVID-19 policies. On a scale between 0 to 100 where 100 = strictest, values ranged from 55.09 - 56.94 in Germany, from 39.35 - 43.06 in Switzerland, and were stable at 37.96 in Austria.

Integrated (Survey Respondent Tweet) Dataset

Survey respondents were prompted for their Twitter handle in order to collect historical tweets about COVID-19 prevention measures.

This is a limitation of the study, as only 79 survey respondents were able to provide historic tweets. The authors attribute this to low Twitter usage in German-speaking countries and account for this limitation by analyzing from a social science perspective.

Survey	Austria	Germany	Switzerland
Participants	565	1,721	274
Twitter Handles	25	77	17

Quantifying Polarization

The authors first analyze for polarization separately in each dataset, measuring dispersion as variance and kurtosis as modality.

A higher variance and lower kurtosis (especially if negative) suggests high levels of polarization.

Additionally, a bimodal coefficient β is used to measure polarization, ranging from 0 to 1. Higher values representing bimodality.

Here, skewness is given as γ, kurtosis as κ, and sample size as n,

$$\beta = \frac{\gamma^2 + 1}{\kappa + 3\frac{(n-1)^2}{(n-2)(n-3)}}$$

Quantifying Polarization

The authors first analyze for polarization separately in each dataset, measuring dispersion as variance and kurtosis as modality.

A higher variance and lower kurtosis (especially if negative) suggests high levels of polarization.

Additionally, a bimodal coefficient β is used to measure polarization, ranging from 0 to 1. Higher values representing bimodality.

Here, skewness is given as γ, kurtosis as κ, and sample size as n,

$$\beta = \frac{\gamma^2 + 1}{\kappa + 3\frac{(n-1)^2}{(n-2)(n-3)}}$$

Quantifying Polarization

The authors first analyze for polarization separately in each dataset, measuring dispersion as variance and kurtosis as modality.

A higher variance and lower kurtosis (especially if negative) suggests high levels of polarization.

Additionally, a bimodal coefficient β is used to measure polarization, ranging from 0 to 1. Higher values representing bimodality.

Here, skewness is given as γ, kurtosis as κ, and sample size as n,

$$\beta = \frac{\gamma^2 + 1}{\kappa + 3\frac{(n-1)^2}{(n-2)(n-3)}}$$

Quantifying Polarization

The authors first analyze for polarization separately in each dataset, measuring dispersion as variance and kurtosis as modality.

A higher variance and lower kurtosis (especially if negative) suggests high levels of polarization.

Additionally, a bimodal coefficient β is used to measure polarization, ranging from 0 to 1. Higher values representing bimodality.

Here, skewness is given as γ, kurtosis as κ, and sample size as n,

$$\beta = \frac{\gamma^2 + 1}{\kappa + 3\frac{(n-1)^2}{(n-2)(n-3)}}$$

On Congruence and the Ecological Fallacy

In the **integrated dataset**, an ecological fallacy would exist if the authors compared correlations in aggregate data to correlations in data of individuals.

The authors mitigate this with human annotation of Twitter accounts, rather than annotating individual tweets. An **agreement score** is assigned to each survey respondent Twitter account to measure congruence of opinions expressed on Twitter with agreement in the survey answers.

This is done using a qualitative content analysis to **inductively categorize Tweet content**. 221 tweets from 20 survey users were categorically labeled, e.g., social and global politics, politicians' handling of pandemic, how dangerous COVID-19 is.

These labels were identical to what was used on the survey and provided an ordinal rating scale of agreement.

The authors calculate the binary inter-annotator agreement, including only perfect matches between survey answers and Twitter annotations.

On Congruence and the Ecological Fallacy

In the **integrated dataset**, an ecological fallacy would exist if the authors compared correlations in aggregate data to correlations in data of individuals.

The authors mitigate this with human annotation of Twitter accounts, rather than annotating individual tweets. An **agreement score** is assigned to each survey respondent Twitter account to measure congruence of opinions expressed on Twitter with agreement in the survey answers.

This is done using a qualitative content analysis to **inductively categorize Tweet content**. 221 tweets from 20 survey users were categorically labeled, e.g., social and global politics, politicians' handling of pandemic, how dangerous COVID-19 is.

These labels were identical to what was used on the survey and provided an ordinal rating scale of agreement.

The authors calculate the binary inter-annotator agreement, including only perfect matches between survey answers and Twitter annotations.

On Congruence and the Ecological Fallacy

In the **integrated dataset**, an ecological fallacy would exist if the authors compared correlations in aggregate data to correlations in data of individuals.

The authors mitigate this with human annotation of Twitter accounts, rather than annotating individual tweets. An **agreement score** is assigned to each survey respondent Twitter account to measure congruence of opinions expressed on Twitter with agreement in the survey answers.

This is done using a qualitative content analysis to **inductively categorize Tweet content**. 221 tweets from 20 survey users were categorically labeled, e.g., social and global politics, politicians' handling of pandemic, how dangerous COVID-19 is.

These labels were identical to what was used on the survey and provided an ordinal rating scale of agreement.

The authors calculate the binary inter-annotator agreement, including only perfect matches between survey answers and Twitter annotations.

On Congruence and the Ecological Fallacy

In the **integrated dataset**, an ecological fallacy would exist if the authors compared correlations in aggregate data to correlations in data of individuals.

The authors mitigate this with human annotation of Twitter accounts, rather than annotating individual tweets. An **agreement score** is assigned to each survey respondent Twitter account to measure congruence of opinions expressed on Twitter with agreement in the survey answers.

This is done using a qualitative content analysis to **inductively categorize Tweet content**. 221 tweets from 20 survey users were categorically labeled, e.g., social and global politics, politicians' handling of pandemic, how dangerous COVID-19 is.

These labels were identical to what was used on the survey and provided an ordinal rating scale of agreement.

The authors calculate the binary inter-annotator agreement, including only perfect matches between survey answers and Twitter annotations.

On Congruence and the Ecological Fallacy

In the **integrated dataset**, an ecological fallacy would exist if the authors compared correlations in aggregate data to correlations in data of individuals.

The authors mitigate this with human annotation of Twitter accounts, rather than annotating individual tweets. An **agreement score** is assigned to each survey respondent Twitter account to measure congruence of opinions expressed on Twitter with agreement in the survey answers.

This is done using a qualitative content analysis to **inductively categorize Tweet content**. 221 tweets from 20 survey users were categorically labeled, e.g., social and global politics, politicians' handling of pandemic, how dangerous COVID-19 is.

These labels were identical to what was used on the survey and provided an ordinal rating scale of agreement.

The authors calculate the binary inter-annotator agreement, including only perfect matches between survey answers and Twitter annotations.

Results

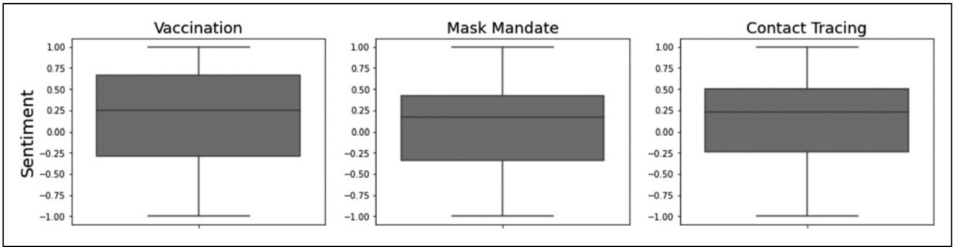

Figure 1. Polarization in Twitter data (all: $n = 90,806$ tweets) in terms of sentiment in the three prevention measures, that is, vaccination, mask wearing, and contact tracing. The sentiments are measured per tweet on a range from -1 for the maximum negative sentiment to $+1$ for the maximum positive sentiment. Tweets with neutral sentiment are excluded. Vaccination shows high variance which indicates a high level of polarization, but also the highest median suggesting a more positive leaning toward the measure.

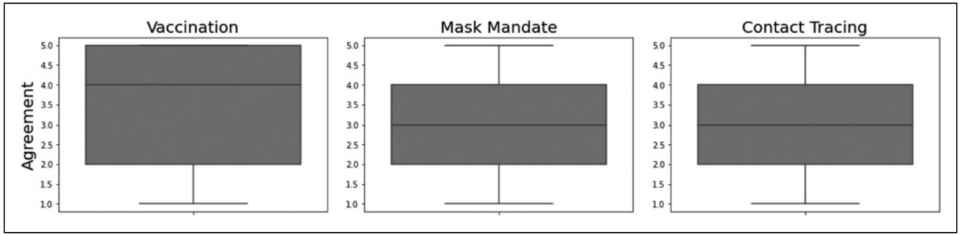

Figure 2. Polarization in Survey data (all: n = 2560 respondents) in terms of agreement to the three prevention measures, that is, vaccination, mask wearing, and contact tracing. The agreement is measured per respondent on a range from 1 for strong disagreement to 5 for strong agreement. Vaccination shows high variance which indicates a high level of polarization, but also the highest median suggesting a more positive leaning toward the measure.

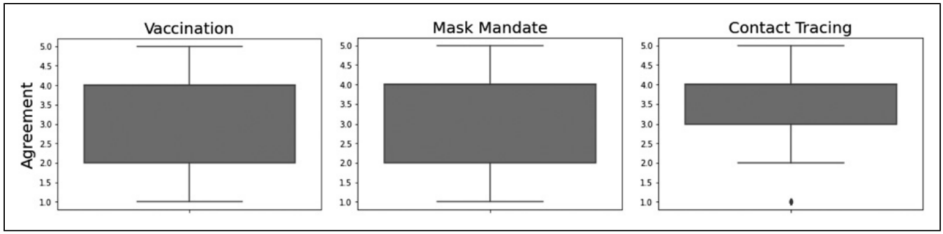

Figure 3. Polarization in the Integrated data (all: n = 79 respondents) in terms of agreement among the three prevention measures, that is, vaccination, mask wearing, and contact tracing. The agreement is measured per respondent on a range from 1 for strong disagreement to 5 for strong agreement. Both, vaccination and mask wearing, show a high variance which indicates a high amount of polarization. All three measures have a median of 4, suggesting a leaning toward approval of the measures.

Overall Findings

The authors report finding polarization to be congruent between Twitter and survey datasets in the measured variables — expressed agreement and extracted sentiment.

They note this to be the first work to consider polarization in both survey and social media data.

Methods Critique

Critique Summary

1. German speakers make relatively light use of Twitter
2. Integrated dataset is small enough that authors appear to switch from quantitative to qualitative approach out of necessity
3. Sentiment analysis is only performed on tweets containing words in the sentiment polarity lexicon
4. Polarization is analyzed only at a point in time, as a state, although the authors note that its temporal dynamics are worth investigating

German-language Twitter usage

WhatsApp is heavily used in Germany.[2] It has been hypothesized that historically high SMS fees catalyzed German WhatsApp adoption.

Surveys have been automatically deployed on that platform[3] using its Business API, which would allow for automatic, temporally-aligned data collection and a wider reach.

[2] Werliin (2020)
[3] Fei et al. (2020)

Limitations of Sentiment Analysis Approach

The sentiment analysis procedure is only performed on tweets containing words linked to the sentiment polarity lexicon.

This excludes 57.37% of the available dataset.

Additionally, the approach used to detect sentiment takes a simple approach to negations, so nuance (i.e., sarcasm) may be associated with the wrong polarity.

Longitudinal Approach

The authors note their future work will involve repeated surveys with identical respondents and questions.

Similarly, automated WhatsApp surveys could be employed here.

Considering the interactive experience of a WhatsApp survey (i.e., as if a chatbot), alternative methods could be employed to capture temporal dynamics of opinion polarization, including factorial vignette approaches requiring shorter quantitative responses.

Discussion

Discussion Questions

1. Considering a temporal/longitudinal approach, what methods could be used to assess relationships between opinion polarization and topics besides COVID-19 prevention measures? Why would this be valuable?
2. What are the limitations of extracted sentiment in this case? Can you think of a "blind spot" of sentiment analysis?

References

Fei, J., Wolff, J., Hotard, M., Ingham, H., Khanna, S., Lawrence, D., ... Hainmueller, J. (2020). *Automated Chat Application Surveys Using WhatsApp*. doi:10.31235/osf.io/j9a2y

Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2022). Polarization of Opinions on COVID-19 Measures: Integrating Twitter and Survey Data. *Social Science Computer Review*, 089443932210876. doi:10.1177/08944393221087662

Werliin, R. (2020, September 17). *New study: Instagram climbs the ladder, TikTok has a long way to go*. AudienceProject. Retrieved September 15, 2022, from https://www.audienceproject.com/blog/key-insights/new-study-instagram-climbs-the-ladder-tiktok-has-a-long-way-to-go/