EasyChair Preprint

№ 6127

Opinion Polarization on COVID-19 Measures: Integrating Surveys and Social Media Data

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

July 21, 2021

7^{th} International Conference on Computational Social Science IC^{2}S^2
July 27-31, 2021, ETH Zürich, Switzerland

Opinion Polarization on COVID-19 Measures: Integrating Surveys and Social Media Data

opinion polarization, surveys, social media, integrating data sources, covid-19 measures

Extended Abstract

Polarization in public opinion is a major issue for societies as high levels can promote adverse effects such as hostility and the spread of misinformation [1]. Research on polarization can be conducted via surveys or social media analyses. One specific type of polarization is opinion polarization that deals with the dispersion of opinions. In survey research, opinion polarization is typically measured by agreement and characterized by the statistics of their distribution [2]. In contrast, social media research uses content-based measures to extract opinions from text and typically analyze them either based on predefined groups along characteristics such as political affiliations, or by derived network-based measures, such as segregation within the topology [3]. Recently, integrating survey and social media data has become an emerging field [4]. However, there is a limitation regarding the comparability of those two research lines and it is not clear whether or not the opinions in surveys and social media content match.

Our research aims to fill this gap by studying polarization from multiple perspectives. We introduce a framework to conduct a comparability analysis that bridges the gap between the offline and online world using an integrated data source. Specifically, we conduct analyses using three types of data sources, i.e., survey, social media, and integrated data that comprises both survey and social media data. We investigate each of the three data sources from two granularity levels: firstly, the fully available data source, and secondly, a comparable subset, e.g., by restricting the survey data to social media users. Our approach uses sentiments for the social media data and agreements for the survey and the integrated data. The congruence between the expressed agreement within the integrated data is verified by manual annotations of a subset of tweets. Moreover, we propose to expand the definition of statistical characteristics to social media analysis to ensure comparability with the opinions expressed in the survey data. Specifically, we apply the bimodality coefficient as a unified measurement that considers the dispersion of data using the skewness γ and excess kurtosis κ. The bimodality coefficient β is defined by the equation $\beta = (\gamma^2+1)/(\kappa+3\frac{(n-1)^2}{(n-2)(n-3)})$, where the sample size n acts as a normalization factor. Each perspective provides a trade-off between data availability, i.e., size of the dataset and presence of social media discussions, and comparability concerning population characteristics and temporal information.

We tested our framework in an analysis of views on the COVID-19 prevention measures in the German-speaking DACH region in the summer of 2020. Our three data sources consist of a representative sample for Internet users of $2,560$ survey respondents, $90,806$ tweets collected from a publicly available Twitter dataset on COVID-19, and a subsample of 79 survey respondents integrated with their social media accounts. Moreover, we use a subset of 705 respondents that use Twitter, $21,479$ tweets from the same period as the online survey, and 20 integrated accounts from which we manually annotated 221 tweets about COVID-19. Thus,

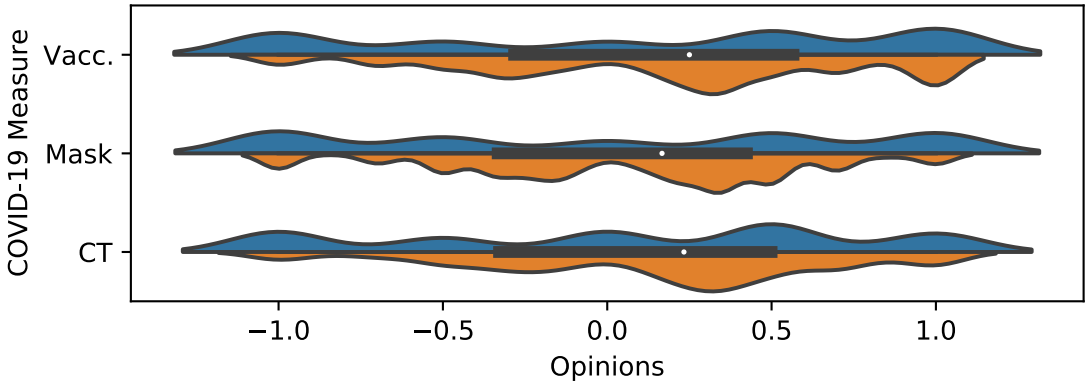

Figure 1: Violinplot of opinions comparing normalized agreement in the full survey dataset (top; blue) to sentiment on the full Twitter dataset (bottom; orange). Vaccination (*Vacc.*) is more polarized compared to *Mask* wearing and contact tracing (*CT*) in both agreement and sentiment due to more extreme opinions skewed towards the positive side. Polarization in agreement is also higher compared to polarization in sentiment.

each perspective retains as much data as possible while enabling cross-perspective comparisons.

We find in all three data sources similar polarization effects (as detailed by opinion dispersion in Figure 1) and a high congruence in the annotated tweets of the integrated data. For instance, vaccination regarding COVID-19 is more polarizing in expressed agreement ($\beta = 0.67$) and sentiment ($\beta = 0.49$) compared to mask-wearing ($\beta = 0.65; 0.44$) and contact tracing ($\beta = 0.59; 0.44$). Moreover, we find that polarization seems to be less prevalent in the subset of Twitter users of the survey respondents compared to the overall survey sample. We suspect that this is due to a bias when respondents consent to data collection, which we aim to investigate in future work. Overall, we conclude that our approach provides a holistic view on polarization of COVID-19 prevention measures in the German language and congruence between the online and offline perspectives.

References

[1] Bessi A, Petroni F, Del Vicario M, Zollo F, Anagnostopoulos A, Scala A, et al. Viral misinformation: The role of homophily and polarization. In: Proceedings of the 24th International Conference on World Wide Web; 2015. p. 355–356.

[2] Bramson A, Grim P, Singer DJ, Berger WJ, Sack G, Fisher S, et al. Understanding Polarization: Meanings, Measures, and Model Evaluation. Philosophy of Science. 2017;84(1):115–159.

[3] Alamsyah A, Adityawarman F. Hybrid sentiment and network analysis of social opinion polarization. In: 2017 5th International Conference on Information and Communication Technology (ICoIC7). IEEE; 2017. p. 1–6.

[4] Stier S, Breuer J, Siegers P, Thorson K. Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. Social Science Computer Review. 2020;38(5):503–516.